



UNIVERSITY OF
LIVERPOOL

Variant Surface Glycoprotein Diversity in African Trypanosomes

Thesis submitted in accordance with the requirements of the
University of Liverpool for the degree of Doctor in Philosophy by

Sara Silva Pereira

May 2018

Author's Declaration

The antibody-based magnetic cell separation optimization was performed by Dr Harry Noyes at the Institute of Integrative Biology of the University of Liverpool. (section 2.3.1).

Tsetse fly infection and feeding for chapter 3 was performed by Aitor Casas-Sánchez and Daniel Southern at the Department of Vector Biology of the Liverpool School of Tropical Medicine.

Protein mass spectrometry in chapter 3 was performed by Dr Dong Xia and Dr Stuart Armstrong at the Institute of Infection and Global Health of the University of Liverpool.

The IL3000 genome used to characterise the VSG expression sites in chapter 4 was sequenced by the Darby laboratory at the Centre for Genomic Research of the University of Liverpool.

The Southern Blot presented in Figure 33 was produced by Dr Simon D'Archivio at the University of Nottingham.

The growth and isolation of *T. vivax* Lins for genome and transcriptome sequencing (chapter 5) was performed by Prof. Rosângela Zacarias Machado at the State University of São Paulo (UNESP), Jaboticabal campus.

Apart from the help and advice acknowledged, this thesis represents the unaided work of the author.

.....

Sara Silva Pereira

15th May 2018

Acknowledgements

First and foremost, I would like to thank my supervisor, Dr Andrew Jackson. Through his continuous guidance and support, he gave me the tools to succeed, whilst making me feel heard and valued. He has been a mentor and a source of inspiration, showing genuine interest and commitment to my development as a researcher and (hopefully) future leader.

A big thanks to Prof. Steve Kemp at the International Livestock Research Institute in Nairobi for access to the ILRI Biobank from which we obtained the historical isolates of *Trypanosoma congolense* and *Trypanosoma vivax*. Also to his office-neighbour, Prof. Eric Fèvre and colleagues, for hosting me during my trips to Kenya.

This thesis would have not been possible without the contributions of Álvaro Acosta-Serrano at the Liverpool School of Tropical Medicine, who provided the fly colonies and of Aitor Casas-Sánchez and Dr Lee Haines, the fly whisperers, who introduced me to the marvellous world of vector biology.

I would also like to acknowledge Dr Matthew Berriman at the Wellcome Sanger Institute for funding part of the historical *T. congolense* sequencing project, and his Parasite Genomics team for much-needed technical support. I must also thank the funders who supported my work: the Bill and Melinda Gates Foundation through the Grand Challenges (Round 11) award, and the BBSRC (BB/M022811/1).

My heartfelt thanks to everyone at the Department of Infection Biology in iC2, for creating the best work environment I could imagine. In particular, my warm gratitude to my colleagues (and now friends for life) Dr Ross Low and Dr Isabel Garcia-Dorival and to the ever-growing Jackson group.

Finally, I would like to thank my parents and my brother for getting me here and not asking too many answerless questions. Last, but absolutely not least, my immense gratitude to my husband, Tiago, for putting up with all my tempers, my late night doubts and my overwhelming enthusiasm for trypanosomes.

Abstract

African trypanosomes are vector-borne haemoparasites that cause animal and human African trypanosomiasis. Survival in the mammal host is mediated by antigenic variation, an immune evasion mechanism based on the sequential replacement of the Variant Surface Glycoprotein (VSG) coat that covers the parasite surface. As the main factors in the host-trypanosome interaction, VSG expression and structure has been extensively studied in *Trypanosoma brucei*. Since trypanosome genomes contain hundreds of highly dynamic VSG genes, studies of VSG diversity are challenging and have been limited to reference strains. However, we expect VSG diversity to be vitally important to disease, both during individual infections and across parasite populations. This thesis examines VSG sequence diversity within and between trypanosome populations to produce new approaches to measuring VSG diversity and a holistic view of antigenic diversity in trypanosome genomes.

In chapters two and five, I present ‘variant antigen profiling’ as a method to characterise VSG repertoires in genomic and transcriptomic data from *T. congolense* and *T. vivax* respectively. Due to species differences in VSG repertoire composition, bespoke methodologies were required. Analysing VSG diversity over space and time was a challenge to conventional techniques that can now be overcome.

In chapter three, I apply variant antigen profiling to metacyclic transcriptomes to characterise *T. congolense* metacyclic VSG expression in the tsetse fly. This revealed that specific phylotypes might be preferentially expressed in infective, metacyclic-stage parasites, which can be taken forward in field experiments.

In chapter four I present a description of putative *T. congolense* expression sites, which are compared to *T. brucei*. In chapter six, I present a comparative analysis of the balance of evolutionary forces affecting molecular evolution of VSGs in *T. brucei*, *T. congolense*, and *T. vivax*. These two chapters bring new insights into antigenic variation evolution in African trypanosomes, aiding the reconstruction of the process of host-parasite coevolution.

Collectively, this work makes a substantial, original contribution to our understanding of antigenic diversity in African trypanosomes, whilst revealing the need for revisiting

basic questions of the mechanisms of antigenic variation in *T. congolense* and *T. vivax*. Such projects will certainly be aided by the multiple applications of variant antigen profiling to gene expression, functional and population studies.

Table of Contents

Author's Declaration	ii
Acknowledgements	iii
Abstract	iv
Table of Contents	vi
List of Figures	xi
List of Tables	xv
List of Abbreviations	xvi
Chapter 1. Introduction	1
1.1 Biology of the parasite	4
1.2 Parasite genetics	7
1.3 Biology of the vector	8
1.4 Disease pathogenesis	10
1.5 Diagnosis	11
1.6 Treatment	15
1.7 Impact	16
1.8 Host-parasite interactions	16
1.9 Antigenic variation	19
1.10 Variant Surface Glycoproteins	20
1.10.1 VSG structure	20
1.10.2 VSG diversity	23
1.10.3 VSG expression	27
1.10.4 VSG switching and recycling	29
1.10.5 Hierarchic activation of VSGs	33
1.10.6 Role of VSGs in pathology and virulence	34
Chapter 2. The Variant Antigen Profile in <i>Trypanosoma congolense</i> : quantifying the frequency of conserved VSG motifs	37
2.1 Introduction	37
2.2 Methods	42
2.2.1 Sample identification and storage	42
2.2.2 Cell lysis and DNA extraction	42

2.2.3 Magnetic antibody cell sorting control	44
2.2.4 Genomic amplification	45
2.2.5 Next-Generation Sequencing (NGS).....	45
2.2.6 Analysis of NGS data	46
2.2.7 Strain variation	51
2.2.8 Statistical analysis	51
2.3 Results	52
2.3.1 Antibody-based cell sorting efficacy	52
2.3.2 <i>De novo</i> assembly.....	54
2.3.3 Sampling test	56
2.3.4 Phylotype universality and phylogeny revision.....	57
2.3.5 Positive Control	59
2.3.6 Sequence similarity vs. structural motif searches	59
2.3.7 Antigenic diversity in <i>T. congolense</i>	62
2.4 Discussion.....	67
2.4.1 Antigenic diversity in <i>T. congolense</i>	67
2.4.2 Strain relationships with space and time	72
2.4.3 The VAP methodology	75
2.4.4 Conclusion	78
 Chapter 3. The metacyclic VSG repertoire of <i>Trypanosoma congolense</i>	
‘savannah’ Tc1/148	79
3.1 Introduction.....	79
3.2 Methods	83
3.2.1 <i>T. congolense</i> Tc1/148 strain	83
3.2.2 Tsetse fly infection and rearing	83
3.2.3 Fly dissection	84
3.2.4 Metacyclic enrichment from fly mouthparts	84
3.2.5 RNA and protein extraction	86
3.2.6 RNA sequencing	86
3.2.7 Sample preparation for proteomics	87
3.2.8 NanoLC MS ESI MS/MS analysis	88
3.2.9 Protein identification and quantification.....	89
3.2.10 Transcriptome profiling.....	89
3.2.11 Statistical analysis	90
3.3 Results	91
3.3.1 Trypanosome populations in the tsetse fly mouthparts	91

3.3.2 Adapting variant antigen profiling to transcriptomic data	94
3.3.3 The expressed mVSG repertoires of trypanosomes	94
3.3.4 The expressed mVSG repertoires of metacyclic-enriched populations ..	101
3.3.5 Further attempts to detect mVSG by MS	105
3.4 Discussion.....	106
3.4.1 Conclusions.....	112
Chapter 4. Characterisation of conserved telomeric structures in <i>T. congolense</i>	
.....	113
4.1 Introduction.....	113
4.2 Methods	117
4.2.1 Parasite stocks and culture	117
4.2.2 DNA extraction	117
4.2.3 Long-read genomic DNA library preparation and sequencing	118
4.2.4 Genome assembly	119
4.2.5 Genome annotation.....	119
4.2.6 Contig selection.....	120
4.2.7 Contig annotation	120
4.2.8 Multiple Sequence Alignment.....	121
4.2.9 Phylogenetic estimation	121
4.2.10 Comparison of tree topology	121
4.2.11 Recombination and selection tests	122
4.3 Results	124
4.3.1 The telomere-containing contigs	124
4.3.2 <i>T. congolense</i> has canonical telomeric structures	125
4.3.3 <i>T. congolense</i> telomere-associated genes lack evidence for sequence adaptation to the telomere	138
4.3.4 Tree Topology	144
4.3.5 The chromosomal location of the telomere-associated structures	145
4.4 Discussion.....	150
4.4.1 Future directions	155
4.4.2 Conclusion	157
Chapter 5. The Variant Antigen Profile to quantify antigen diversity in	
<i>Trypanosoma vivax</i>	158
5.1 Introduction.....	158
5.2 Methods	163
5.2.1 Sample identification and storage	163

5.2.2 Cell lysis and nucleic acid extraction.....	164
5.2.3 Next-Generation Sequencing (NGS).....	165
5.2.4 VSG-like sequence recovery.....	166
5.2.5 Multiple Sequence Alignment.....	168
5.2.6 Phylogenetic estimation	168
5.2.7 Clusters of Orthologous Groups (COGs) identification	168
5.2.8 Binary matrices	169
5.2.9 Strain variation	169
5.2.10 <i>T. vivax</i> Lins transcriptome analysis	170
5.2.11 Ethics Statement.....	170
5.3 Results	171
5.3.1 Genome completion of sequenced field strains	171
5.3.2 Sampling test	172
5.3.3 Intra-phylo type variation.....	176
5.3.4 Genetic relationship between strains	179
5.3.5 The genomic VSG repertoire of <i>T. vivax</i> Lins	181
5.3.6 The expressed VSGs in <i>T. vivax</i> Lins	186
5.4 Discussion.....	190
5.4.1 Antigenic diversity in <i>T. vivax</i>	191
5.4.2 The COG matrix methodology	194
5.4.3 Conclusion and Future directions.....	196
Chapter 6. The molecular evolution of VSGs in African trypanosomes.....	197
6.1 Introduction.....	197
6.2 Methods	201
6.2.1 Genomes.....	201
6.2.2 VSG mapping.....	202
6.2.3 VSG characterisation	202
6.2.4 VSG donor analysis	204
6.2.5 Recombination analysis	205
6.3 Results.....	206
6.3.1 VSG mapping.....	206
6.3.2 VSG characterisation	207
6.3.3 Donor analysis	212
6.4 Discussion.....	218
6.4.1 Conclusion	223
Chapter 7. General Discussion.....	224

7.1 Conclusion	232
References	233

List of Figures

Figure 1 Life cycle of <i>T. brucei</i> , <i>T. congolense</i> , and <i>T. vivax</i>	6
Figure 2 The anatomy of the tsetse fly (<i>Glossina sp.</i>).....	10
Figure 3 Sample collection in microhaematocrit capillaries for assessment of blood PCV and examination of wet preparations of the buffy coat. tic test (CEVA, France), which detects <i>T. vivax</i> and <i>T. congolense</i> , separately or in a multiple infection of the same host (Pillay et al. 2013).	14
Figure 5 Host-Parasite interactions: ‘waves of parasitaemia’ in a trypanosome infection.....	20
Figure 6 The VSG structure. a) The 3D model of VSG221 dimer from <i>T. brucei</i>	21
Figure 7 Phylogenetic tree topologies for VSG-like subfamilies in <i>T. brucei</i> , <i>T. congolense</i> and <i>T. vivax</i>).	26
Figure 8 Mechanisms of VSG switching.	32
Figure 9 C terminal domains (CTD) of <i>T. congolense</i> IL3000 VSGs	41
Figure 10 Diagnostic motif consensus sequences of 15 <i>T. congolense</i> VSG phylotypes and their positioning in the primary amino acid structure (1 to 450 amino acids).....	50
Figure 11 Fluorescence-activated cell sorting of CD45+ cells.	53
Figure 12 Gel electrophoresis UV picture of PCR products from CD45+ cell depletion.....	54
Figure 13 Comparison of performance de <i>nov</i> o assemblers. Velvet, Abyss and SOAPdenovo2 were compared, using N50, contig number and maximum length of contigs as measured in three different <i>T. congolense</i> strains.....	55
Figure 14 Assembly quality comparison with different kmer values, using Velvet 1.2.10 (Zerbino 2010).....	56
Figure 15 Maximum likelihood phylogeny of <i>T. congolense</i> full-length VSG using IL3000 (Kenya), IL3674 (The Gambia), and IL3900 (forest sub-type).	59
Figure 16 Performance of the protein motif-based VAP compared to the manual estimation of phylotype proportion in IL3000 and all the field isolates.	61
Figure 17 The relationships between the VSG repertoire, geography and population structure in <i>Trypanosoma congolense</i>	63
Figure 18 Variation in phylotype proportions estimated from the real (left) and the randomized (right) data (N = 41 for each condition).....	64
Figure 19 Phylotype variation across the population.	66

Figure 20 The relationships between the VSG repertoire, geography and population structure in <i>Trypanosoma congolense</i> including the isolates published by Tihon <i>et al.</i> (2017).	73
Figure 21 <i>T. congolense</i> in the tsetse fly hypopharynx.....	92
Figure 22 Immunofluorescence detection of trypanosome populations from the tsetse fly mouthparts on a confocal microscope.....	93
Figure 23 Transcriptomic Variant Antigen Profiles of trypanosomes extracted from tsetse mouthparts.....	96
Figure 24 Comparison of average phylotype relative abundance (adjusted for transcript abundance) in transcriptomic samples and genomic profiles from a random selection of VSGs of Tc1/148 (mean \pm σ).	98
Figure 25 VSG transcript abundances.....	99
Figure 26 Maximum likelihood phylogeny of phylotype 8 estimated from protein sequences.....	100
Figure 27 <i>Trypanosoma</i> populations before and after DE52-cellulose column separation (mean \pm σ).	102
Figure 28 Transcriptomic Variant Antigen Profiles of trypanosomes extracted from pooled tsetse mouthparts after metacyclic-parasite enrichment.	103
Figure 29 The consensus structure of the <i>T. brucei</i> Lister 427 bloodstream expression site (BES).....	115
Figure 30 The genomic context of the <i>T. congolense</i> IL3000 actively-transcribed VSG, or the 'Active VSG ES'.....	116
Figure 31 The consensus structure of the telomere-associated structures in <i>T. congolense</i>	125
Figure 32 Multiple sequence alignment of the 369 bp repeat in IL3000 and Tc1/148.	128
Figure 33 The karyotype of <i>T. brucei</i> spp. and <i>T. congolense</i>	129
Figure 34 CNR1 sequence variation and frequency in Tc1/148.	130
Figure 35 Multiple sequence nucleotide alignment of CNR2.	133
Figure 36 Multiple sequence nucleotide alignment of CNR3.	135
Figure 37 Multiple sequence nucleotide alignment of CNR4.	137
Figure 38 Consensus maximum likelihood phylogeny of the transferrin receptors protein sequences from <i>Trypanosoma congolense</i> and <i>Trypanosoma brucei</i>	139
Figure 39 Consensus maximum likelihood phylogeny of Fam53 protein sequences from African Trypanosomes.	140

Figure 40 Consensus maximum likelihood phylogeny of cathepsin B protein sequences from African Trypanosomes.....	141
Figure 41 Consensus maximum likelihood phylogeny of DEAH-box RNA helicase protein sequences from <i>Trypanosoma congolense</i> and <i>Trypanosoma brucei</i>	143
Figure 42 The difference in phylogenetic signal along the conserved non-coding regions of the ES.....	145
Figure 43 Structure of the telomere-containing contigs from the Tc1/148 megabase chromosomes.....	147
Figure 44 The non-redundant structure of the <i>T. congolense</i> mini-chromosomes in Tc1/148 and IL3000.	149
Figure 45 Methodology followed for the <i>T. vivax</i> VAP development.....	167
Figure 46 Proportion of VSG belonging to families 23-26 from all field strains compared to the proportions of the reference full VSG repertoire given as mean $\pm \sigma$	175
Figure 47 Phylogenies of the VSG families.....	176
Figure 48 The relationship between <i>T. vivax</i> SNPs and COGs.	180
Figure 49 Heatmap and cladogram showing presence and absence of 1081 VSG orthologues across our sample dataset.	182
Figure 50 Maximum likelihood phylogeny of <i>T. vivax</i> strains based on whole genome SNPs estimated with GATK (N = 1,011,378).	183
Figure 51 Relative Frequency of each VSG family in our strain dataset, TvY486, and TvLins.....	184
Figure 52 Maximum likelihood phylogeny of amino acid sequences of VSG families 23-26 using WAG+F model of amino acid substitution.	185
Figure 53 Correlation of transcript abundances based on the 1000 most abundant transcripts between S3, S4, and IL1392. between the ratio of FPKM values of S3/S4 and IL1392 respectively.	188
Figure 54 FPKM values of top 6 transcripts and correspondent VSG families for S3.	189
Figure 55 Reference VSG mapping strategy.	202
Figure 56 VSG characterisation strategy.	203
Figure 57 VSG donor analysis strategy.	204
Figure 58 Proportion of field strain paired VSG reads remaining paired when mapped to full-length VSG for each African trypanosome species.....	207
Figure 59 Example representation of fully coupled and multi-coupled VSGs with their respective topology probability.....	208

Figure 60 The composition of the VSG repertoire for each African trypanosome species (mean \pm SEM).	209
Figure 61 Phylogenetic incompatibility among VSG genes using PhiPack (Bruen et al. 2006).	211
Figure 62 Donor VSG sequence localisation within <i>T. congolense</i> uncoupled VSGs.	212
Figure 63 Nucleotide sequence identity amongst all sequences involved in MC VSG formation (donors and recipients) (mean \pm 95 % CI).	214
Figure 64 <i>T. congolense</i> VSG recombination distribution by phylotype.	214
Figure 65 The sequence coverage of VSG donors expressed as a percentage of the full multi-coupled VSG repertoire.	215
Figure 66 The sequence coverage of VSGs with a single donor, represented as a percentage of VSG.	216
Figure 67 Total sequence orthology amongst VSG repertoires of the same species.	217

List of Tables

Table 1 Species of African trypanosomes. Details compiled from Gibson (2007) and Hutchinson & Gibson (2015).	3
Table 2 The surface of African trypanosomes according to the Cell Surface Phylome (Jackson et al. 2013).	18
Table 3 Summary of VSG-like genes in African trypanosomes (Jackson et al. 2012).	24
Table 4 <i>T. congolense</i> strains used in this study.	44
Table 5 Correlation analysis results (partial vs. full repertoire VAP). r = Pearson's moment correlation.	57
Table 6 Correlation values for individual phylotypes before and after improvement (Pearson's Moment Correlation) and description of the outliers removed.	60
Table 7 Sequencing statistics and number and expression values of VSG transcripts recovered per transcriptome.	95
Table 8 Proteomic Results from mass spectrometry sequencing of metacyclic-enriched parasite suspensions obtained by DE52-cellulose separation.	104
Table 9 Variant antigen profiles of the genomic VSG and the telomeric VSG (ES) from IL3000 and Tc1/148.	131
Table 10 Hypothetical proteins found inside the ES of Tc1/148 and their available expression data.	144
Table 11 Sample ID, date and location of collection, host, and species used for passaging.	164
Table 12 Sequencing coverage and mapping results of field strains compared to the reference strain.	172
Table 13 Number of VSG recovered from each strain by sequence similarity search and their proportion compared to the reference full repertoire.	174
Table 14 Distribution of COGs recovered by the two approaches under different sensitivity thresholds.	177
Table 15 Summary of VSG transcripts in S3 and S4 compared to the IL1392 transcriptome previously published by Jackson et al. (2015).	186
Table 16 Description of samples used in this chapter.	201

List of Abbreviations

A.	<i>Anaplasma</i>
AAT	<i>Animal African Trypanosomiasis</i>
AC	Adenylate Cyclases
ACK	Ammonium-Chloride-Potassium
AIC	Akaike Information Criterion
APOL-1	Apolipoprotein L1
BAC	Bacterial-Associated Chromosome
BES	Bloodstream Expression Site
BI	Bayesian Inference Phylogeny
BLAST	Basic Local Alignment Search Tool
bp	Base Pairs
BSA	Bovine Serum Albumin
BSF	Bloodstream Form
CATT	Card Agglutination Test
CGR	Centre For Genomic Research
chIP-sequencing	Chromatin Immunoprecipitation Sequencing
CI	Confidence Interval
CNR	Conserved Non-Coding Region
COG	Cluster Of VSG Orthologs
CSA	Glycosaminoglycan Chondroitin Sulphate A
CTD	C-Terminal Domain
DBL α	Duffy Binding Like Alpha
DES	Dominant Expression Site
DNA	Deoxyribonucleic Acid
dNTP	Deoxynucleotide
DSB	Double-Strand Breaks
DTM	Differentiating Trypanosome Medium
E-value	Exponent Value (In BLAST And HMM Search)
EDTA	Ethylenediaminetetraacetic Acid
EF	Epimastigote Forms
ELISA	Enzyme Linked Immunosorbent Assay
ENA	European Nucleotide Archive
ES	Expression Site
ESAG	Expression Site Associated Gene
ESB	Expression Site Body

EST	Expression Sequence Tag
FC	Fully Coupled
FEL	Fixed Effects Likelihood
FPKM	Fragments Per Kilobase Of Transcript Per Million Mapped Reads
FUBAR	Fast Unbiased Bayesian Approximation
G.	<i>Glossina</i>
Γ or G	Gamma Correction
GARD	Genetic Algorithm For Recombination Detection
GATK	Genome Analysis Toolkit Suite
GPI	Glycosylphosphatidylinositol
GRESAG	Gene Related ESAG
HAT	Human African Trypanosomiasis
HGAP	Hierarchical Genome Assembly Process
HMM	Hidden Markov Model
IFAT	Indirect Fluorescent Antibody Test
IFN- γ	Interferon Gamma
IgG	Immunoglobulin G
IgM	Immunoglobulin M
IL	Interleukin
ILRI	International Livestock Research Institute
ITS	Internal Transcribed Spacer
JTT	Jones, Taylor And Thornton Model Of Amino Acid Substitution
Kb	Kilobase Pairs
KH test	Kishinoi-Hasegawa Test
LM	Light Microscope
InL	Log-Likelihood
MHC-II	Major Histocompatibility Complex
MASP	Mucin-Associated Surface Protein
MC	Multi-Coupled
MCMC	Markov Chain Monte-Carlo
MEM	Minimum Essential Medium
MES	Metacyclic Expression Site
MF	Metacyclic Forms
ML	Maximum Likelihood Phylogeny
MS	Mass Spectrometry
mSP	Major Surface Protein

mVSG	Metacyclic VSG
NGS	Next-Generation Sequencing
NJ	Neighbor-Joining Phylogeny
NLP	Nucleoplasmin-Like Protein
NTD	N-Terminal Domain
p-value	Probability Value
<i>P.</i>	<i>Plasmodium</i>
PacBio	Pacific Biosciences
PAG	Procyclic-Associated Gene
PBS	Phosphate Buffered Saline
PCA	Principal Component Analysis
PCF	Procyclic Form
PCR	Polymerase Chain Reaction
PCV	Packed Cell Volume
pfEMP1	Plasmodium Falciparum Erythrocyte Membrane Protein-1
PFGE	Pulsed Field Gel Electrophoresis
PHI	Pairwise Homoplasy Index
PI	Phylogenetic Incompatibility
Pol I	Polymerase I
Pol II	Polymerase II
rDNA	Ribosomal DNA
REL	Random Effects Likelihood
RFLP	Restriction Fragment Length Polymorphism
RNA	Ribonucleic Acid
RT-PCR	Real-Time PCR
σ	Standard Deviation
SEM	Standard Error Of The Meam
SGC	Segmental Gene Conversion
SMRT	Single-Molecule Real-Time
SMS	Smart Model Selection
SNP	Single Nucleotide Polymorphism
SRA	Serum Resistance-Associated Gene
SSU rRNA	Small Subunit Ribosomal Ribonucleic Acid
<i>T.</i>	<i>Trypanosoma</i>
TAR	Transformation-Associated Recombination
Hp-Hb	Haptoglobin-Hemoglobin

Tc1/148	T. congolense 'savannah' 1/148 (Mboi/Ng/60/1-148)
TFR	Transferrin Receptors
TgsGP	T. B. Gambiense-Specific Glycoprotein
TLF-1/-2	Trypanolytic Factors 1 And 2
TNF- α	Tumour-Necrosis Factor Alpha
UC	Uncoupled
UM	Unmapped
UTR	Untranslated Region
VAP	Variant Antigen Profiling
VAT	Variable Antigen Type
VESA-1	Variant Erythrocyte Surface Antigen-1
VR	VSG-Related
VSG	Variant Surface Glycoprotein
WAG	Whelan And Goldman Model Of Amino Acid Substitution
WHO	World Health Organisation
YAC	Yeast-Associated Chromosome

Chapter 1. Introduction

The Kinetoplastida (Euglenozoa) are unicellular flagellates that include the free-living Bodonids and the parasitic trypanosomatids. In addition to the nuclear genome, these organisms have defining characteristics, such as the kinetoplast, a mitochondrial-derived genome, and a glycosome, a spherical organelle with a protein-dense matrix but no genome (Lopes 2010). Trypanosomatids include the genera *Trypanosoma* and *Leishmania*.

Leishmania parasites cause leishmaniasis, a disease with worldwide distribution that can cause visceral leishmaniasis, in which parasites invade and proliferate in liver, spleen and bone marrow of the host; cutaneous leishmaniasis, which is characterised by a localized long-term chancre at the bite site; and mucocutaneous leishmaniasis, which progresses from the cutaneous form to chronic destruction of mucosal tissue (Peacock et al. 2008).

The genus *Trypanosoma* includes hundreds of different species with a vast vertebrate and invertebrate host range. The stercorarian (American) and salivarian (African) trypanosomes are clinically relevant as the causes of Chagas disease, sleeping sickness, nagana, and surra (Lopes 2010).

Chagas disease, caused by *T. cruzi* and transmitted by triatomine bugs (*Triatominae* sp.), is characterised by an acute phase that can be lethal, and a chronic, long-lasting phase, defined by cardiomyopathy or digestive mega-syndromes, and ultimately death (Stuart et al. 2008).

Sleeping sickness, or human African trypanosomiasis (HAT), is a disease transmitted by tsetse flies (*Glossina* sp.) and caused by *T. brucei rhodesiense* in East Africa and *T. brucei gambiense* in West Africa (Malvy & Chappuis 2011).

Nagana, or animal African trypanosomiasis (AAT), is a wasting disease of livestock, cyclically transmitted by tsetse flies in sub-Saharan Africa, and mechanically transmitted by other haematophagous flies in Africa and South America (Giordani et al. 2016). It is caused by *Trypanosoma brucei brucei*, *Trypanosoma congolense*, and *Trypanosoma vivax*.

Dourine and surra are diseases with worldwide distribution, caused by *Trypanosoma brucei equiperdum* and *Trypanosoma brucei evansi*, respectively (Brun et al. 1998).

In this thesis, I will focus on African trypanosomes. African trypanosomes are extracellular protozoan parasites of the subgenera *Trypanozoon*, *Nannomonas*, *Duttonella*, and *Pycnomonas*, which cause severe wasting diseases in a range of mammal hosts. Taxonomy of African trypanosomes has been a subject of debate, but at present there are eight species identified (**Table 1**). Veterinary relevant species include *T. (Trypanozoon) brucei* spp., *T. brucei evansi*, *T. brucei equiperdum*, *T. (Nannomonas) congolense*, *T. (Duttonella) vivax*, and *T. (Pycnomonas) suis*. Medically important species are *T. brucei gambiense* and the zoonotic agent *T. brucei rhodesiense*, both causes of human sleeping sickness.

Table 1 Species of African trypanosomes. Details compiled from Gibson (2007) and Hutchinson & Gibson (2015).

Subgenus	Species	Subspecies	Transmission	Host range	Distribution
<i>Trypanozoon</i>	<i>T. brucei</i>	<i>T. b. brucei</i>	Cyclical	Wild & domestic animals	Tsetse belt
		<i>T. b. rhodesiense</i>	Cyclical	Humans, wild & domestic animals	East Africa
		<i>T. b. gambiense</i>	Cyclical	Humans	West Africa
	<i>T. equiperdum</i>	-	Sexual	Equids	Worldwide
	<i>T. evansi</i>	-	Mechanical	Wild & domestic animals	Asia, North Africa, Middle East, Central and South America
<i>Nannomonas</i>	<i>T. congolense</i>	<i>T. c. forest</i>	Cyclical	Wild & domestic animals	Tsetse belt
		<i>T. c. Kilifi</i>	Cyclical	Wild & domestic animals	Tsetse belt
		<i>T. c. savannah</i>	Cyclical	Wild & domestic animals	Tsetse belt
	<i>T. simiae</i>	-	Cyclical	Wild & domestic animals	Tsetse belt
		<i>T. s. tsavo</i>	Cyclical	Suids	
	<i>T. godfreyi</i>	-	Cyclical	Wild & domestic animals	Tsetse belt
<i>Duttonella</i>	<i>T. vivax</i>	-	Cyclical & mechanical	Wild & domestic animals	Tsetse belt & South America
<i>Pycnomonas</i>	<i>T. suis</i>	-	Cyclical	Suids	Tsetse belt

The majority of African trypanosomes are transmitted by tsetse flies (*Glossina sp.*). Therefore, most biological transmission of AAT is restricted to the tsetse belt, the endemic region of tsetse flies, spanning from the Sub-Saharan region to the Kalahari Desert. However, *T. vivax*, *T. brucei evansi*, and *T. brucei equiperdum* have evolved tsetse-independent life cycles. *T. vivax* and *T. brucei evansi* are transmitted by tabanids (*Tabanidae*) (Hoare 1972) and stable flies (*Stomoxys spp.*) (Levine 1973), which allowed the spread of both species into Latin America (Osório et al. 2008) and of *T. brucei evansi* to North Africa, Middle East, and Asia. Mechanical transmission of these two species also exists within the tsetse belt (Desquesnes & Dia 2003; Desquesnes & Dia 2004; Mossaad et al. 2017). *T. brucei evansi* has other transmission modes that include mechanical transmission by bats, and vertical, horizontal, iatrogenic, and per-oral transmission (Brun et al. 1998; Desquesnes et al. 2013). *T. brucei equiperdum* is sexually or vertically transmitted and therefore can spread worldwide (Brun et al. 1998). With the exception of *T. brucei equiperdum*, vertical transmission of AAT through transplacental transmission is of epidemiological relevance in South America only, having been documented in Venezuela and Brazil (Ogwu & Nuru 1981; Osório et al. 2008).

1.1 Biology of the parasite

When a tsetse fly is infected with *T. brucei*, stumpy bloodstream forms colonise the insect gut, where they lose their variant surface glycoprotein (VSG) coat and differentiate into rapidly dividing procyclic forms, expressing procyclin in their cell surface. The parasites migrate anteriorly towards the salivary glands, where they differentiate into epimastigotes attached to the epithelial layer, and subsequently into small, free-swimming infective metacyclics, which express metacyclic VSG (mVSG). This VSG coat is the first line of defence against the initial immune response of the mammal host (Turner et al. 1988). The differentiation from epimastigotes to metacyclics often involves intermediate stages. When the fly feeds on a mammal, the parasites are injected with the saliva into the bloodstream, where they develop to slender bloodstream forms within the first 10 days of infection. As the bloodstream infection evolves, a proportion of slender bloodstream forms transforms into non-dividing short stumpy forms, which make the link between the host and the vector infections (reviewed in Macgregor & Matthews 2010). Human-infective *T. brucei* often invades the central nervous system (Malvy & Chappuis 2011). Recently, *T. brucei* has been reported in the skin and adipose tissue of the mammal host

(Capewell et al. 2016; Trindade et al. 2016; Tanowitz et al. 2017), and such dissemination may be under-represented. The greatest developmental differences between African trypanosomes are observed in the vector stages (Hoare 1972).

T. congolense has a similar life cycle to *T. brucei*, although differentiation into epimastigotes starts in the proventriculus, followed by migration to the proboscis and hypopharynx, where they become infective metacyclics, bypassing the salivary glands (Peacock et al. 2012). There is no morphological change between slender and stumpy forms (**Figure 1**), although at peak parasitaemia, *T. congolense* also reduces the number of proliferative forms (Silvester et al. 2017). In both species, bottleneck selection occurs during migration from the midgut to the salivary glands or proboscis.

T. vivax lacks a procyclic stage. Following a short colonisation of the proventriculus and foregut as long trypomastigotes, *T. vivax* migrates to the tsetse mouthparts, where it differentiates into an epimastigote in the cibarium and proboscis and then into infective metacyclic trypomastigotes in the proboscis and hypopharynx (Ooi et al. 2016). From the hypopharynx, it can be injected into a new mammal host at the next bloodmeal. In the mammal host, *T. vivax* has also been reported outside the vascular system, particularly in the heart, lymphatic and central nervous system tissues, causing neurological symptoms and sporadic acute haemorrhagic infections (Magona et al. 2008).

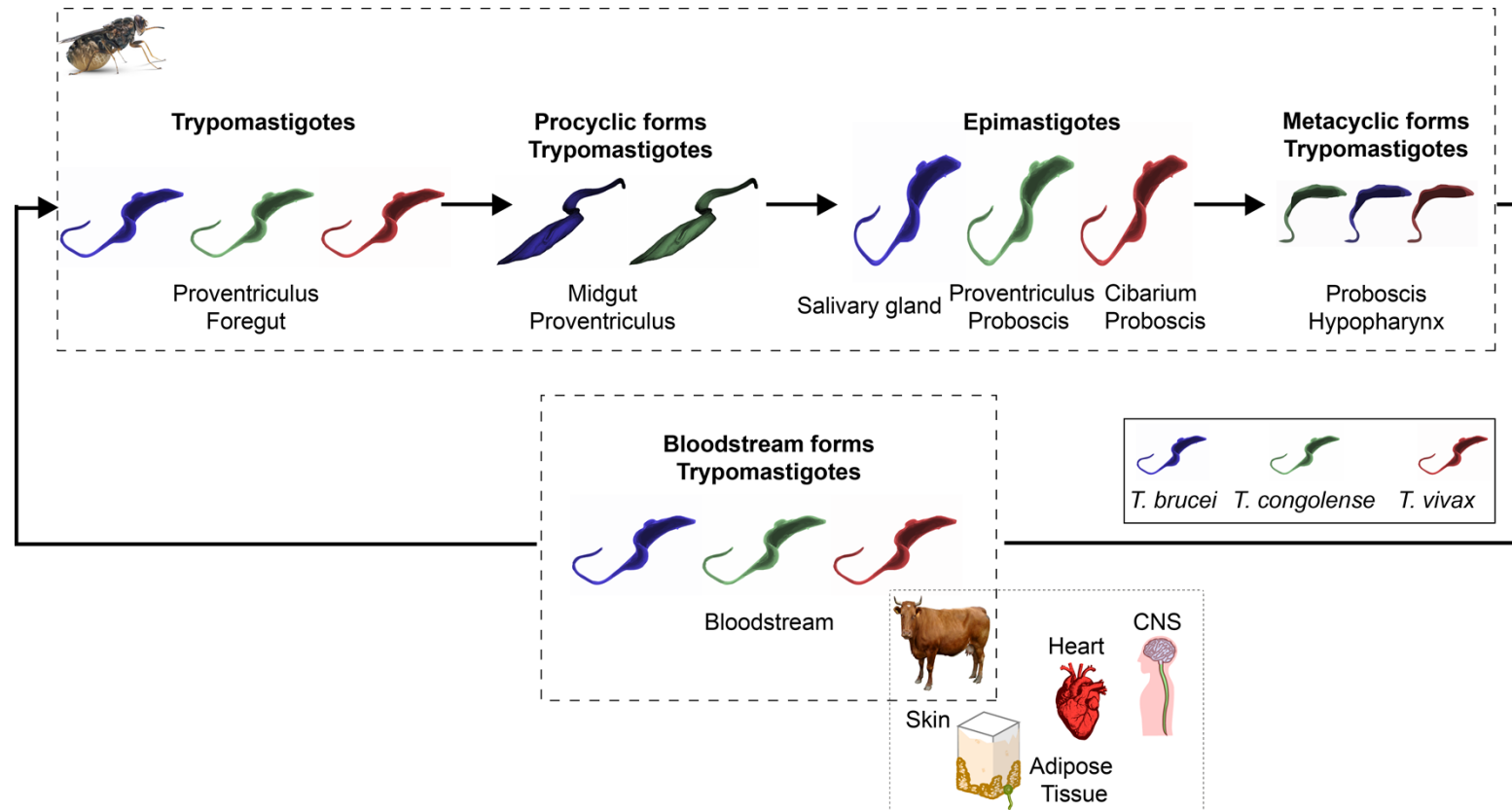


Figure 1 Life cycle of *T. brucei*, *T. congolense*, and *T. vivax*. A tsetse fly takes a blood meal on an infected mammal and becomes a vector of trypanosomiasis. Procyclics establish in the midgut where clonal expansion occurs. The parasites travel to the proventriculus, salivary glands and/or proboscis, where they become epimastigotes and then infective metacyclics. In the following bloodmeal, the fly transmits a portion of these parasites in the saliva to the mammal host. The parasites relocate to the bloodstream as metacyclic trypomastigotes and differentiate to bloodstream forms. Parasites may be found in the central nervous system (*T. brucei*, *T. vivax*), skin (*T. brucei*), adipose tissue (*T. brucei*), and heart (*T. brucei*, *T. vivax*).

1.2 Parasite genetics

African trypanosomes are diploid organisms, whose karyotype has been historically challenging to clarify due to the absence of chromosome condensation during cell division (Vickerman & Preston 1970). The development of pulsed field gel electrophoresis (PFGE) as a means to separate chromosomes based on size and genome sequencing have allowed great improvements in the field.

The *T. brucei* TREU927 genome, which is the reference but one of the smallest *T. brucei* genomes analysed to date, contains 11 megabase chromosomes and a large, but variable number of small and intermediate chromosomes of 30 to 700 kb in size, which have similar sequences to those in the subtelomeres of megabase chromosomes (Melville et al. 1998; Berriman et al. 2005). The subtelomeres, which are the regions between the telomere and the first housekeeping genes, are unusually long in African trypanosomes (Callejas et al. 2006). They can account for up to 75 % of the chromosome length due to copy number polymorphisms derived from divergence of conserved multi-copy gene families (Callejas et al. 2006). In *T. brucei*, they encode more than 20 % of the genes, the majority of which are species-specific and linked to the mechanism of antigenic variation (Berriman et al. 2005). The megabase chromosomes are organised into long non-overlapping gene clusters, which are transcribed as polycistrons and undergo trans-splicing and polyadenylation (Berriman et al. 2005). Coding sequences account for 50 % of the nuclear genome of *T. brucei*.

The genome was sequenced by a combination of whole chromosome shotgun and bacterial artificial chromosome walking strategies. Therefore, most subtelomeric regions, the intermediate and small chromosomes were not included in the genome (Berriman et al. 2005). At publication, it contained 9068 genes and 904 pseudogenes, with 50.9 % GC content. Of these coding sequences, 806 were VSGs. These numbers were vastly underestimated. In fact, further analysis revealed that the subtelomeres are three times larger than the core, mostly comprised of VSG arrays (Callejas et al. 2006). The genome assembly is actively being curated, currently including 1482 VSG-like sequences, 1313 (89 %) of which are pseudogenic or degenerate sequences

(<http://www.genedb.org/Homepage/Tbruceibrucei927>). The *T. brucei* Lister 427 genome is much larger than TREU927, particularly due to differences in the size of

the subtelomeres between homologous chromosomes (Melville et al. 2000). This results in a larger VSG pool of 2563 sequences (Cross et al. 2014), indicating that the size of the megabase chromosomes and the VSG repertoire can vary between strains.

The *T. congolense* genome shares the karyotype characteristics of the *T. brucei* genome. The cell surface related genes account for the main differences in genomic content (Jackson et al. 2012). The same applies to *T. vivax*, with the exception that this species only has one or two mini-chromosomes (Dickin & Gibson 1989). This might potentially impact on the capacity for antigenic variation because, in *T. brucei* and *T. congolense*, the mini-chromosomes are rich in VSGs (Wickstead et al. 2004).

1.3 Biology of the vector

Tsetse flies include thirty-one species of the genus *Glossina*, usually classified into three sub-genera: *Nemorhina* (palpalis group) common in vegetated areas near water in Western and central Africa; *Glossina* sensu stricto (morsitans group) found in woodland 'savannah' where animals are common; and *Austenina* (fusca group), endemic in forest belts and therefore endangered by increased human activity in these areas. Although all tsetse species can potentially be biological vectors of African trypanosomiasis, *G. fuscipes*, *G. palpalis*, and *G. morsitans* (subgroups palpalis and morsitans) are the most relevant in nature (Franco et al. 2014).

Tsetse flies require specific conditions of temperature (16°C–38°C) and humidity (50 %–80 %) and thus prefer regions with water bodies and dense vegetation, avoiding direct sunlight and dry wind. These conditions are also ideal for most terrestrial mammals where tsetse flies feed, making it optimal foci for African trypanosomiasis.

Tsetse flies are robust flies (6-14mm) that can live up to four months (**Figure 2**). They are holometabolous and viviparous, as the larvae develop from the egg in the uterus of the female fly, where they progress through all three larval stages. The female fly gives birth to a single adult larva, which burrows into the soil where it pupates in the first five hours. Adult flies emerge from pupal stage after three weeks. The teneral adult fly searches for its first bloodmeal, a point where is most susceptible to trypanosomal infection. Maturation, mating and first birth occur within 12-14 days, and then a single larva will be produced every 9 days for the rest of the

fly life. Tsetse flies have a slow rate of reproduction, which would facilitate vector control measures, however, fly populations are resilient and persist at very small levels. Their fast movement (20km/hour, 2x30min everyday) also promotes fast re-infestation by neighbouring populations after local eradication.

Trypanosome species have divergent developmental cycles in tsetse flies. *T. brucei* and *T. congolense* first colonise the fly midgut and subsequently travel anteriorly, crossing the proventriculus towards the salivary glands and mouthparts respectively, where they differentiate into infective metacyclics. Although *T. vivax* can be found as posteriorly as the foregut and proventriculus during the first day of infection (Ooi et al. 2016), it is mostly restricted to the proboscis. This feature allows mechanical transmission by non-tsetse flies.

Fly infection rates in the field are about 0.1 % for *T. brucei*, 2 % for *T. congolense* and 10-20 % for *T. vivax* (Haines 2013), which are consistent with the susceptibility of adult flies to become infected with trypanosomes: *T. brucei* mostly infects teneral flies during their first blood meal, *T. congolense* preferentially affects teneral flies, but can infect later on, whereas *T. vivax* is thought to infect flies of any age (Haines 2013). Transmission of AT to the mammal host is dependent on number of trypanosomes inoculated at the bloodmeal, but a single bite from an infected fly can transmit the infection to a naïve mammal host (Thuita et al, 2008). Therefore, disease transmission is affected by trypanosome species, vector and host susceptibility to infection, density of fly population, fly age at infection, and frequency of host-vector contacts (Roditi & Lehane 2008; Haines 2013).

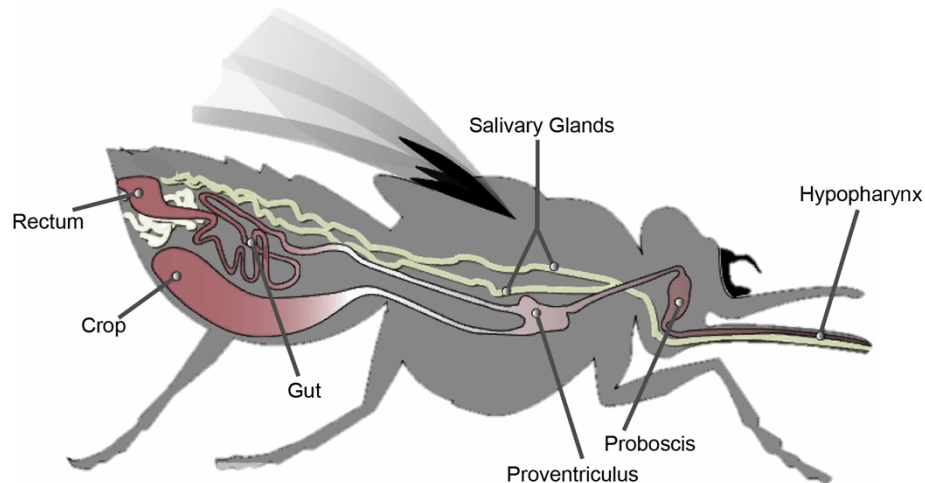


Figure 2 The anatomy of the tsetse fly (*Glossina sp.*). Cartoon shows the positioning of key organs of the tsetse fly. The hypopharynx and proboscis, the proventriculus, the salivary glands, and the gut are colonised by African trypanosomes at different points of the parasites' life cycle. Adapted from ITM (2016).

1.4 Disease pathogenesis

HAT is a two-stage disease. The first stage, 'the haemolymphatic stage', is characterised by parasites colonising the vascular system, which causes fever, malaise, pruritus, lymphadenopathies, and oedema of the face and/or extremities. Sporadically, myocarditis, splenomegaly and hepatomegaly may occur (Malvy & Chappuis 2011). In the second stage, "the meningo-encephalitic stage", parasites invade the central nervous system, causing neurological and endocrinal symptoms, and ultimately leading to death (Malvy & Chappuis 2011). Specific symptoms include circadian rhythm dysregulation, confusion, tremor, involuntary muscle twitching, weakness, unilateral paresis, akinesia or dyskinesia, diffuse hyperpathia, speech impairment, thyroid and adrenocortical dysregulation, and progressive dementia (Malvy & Chappuis 2011).

T. b. gambiense and *T. b. rhodesiense* are morphologically indistinguishable, but the dynamics of disease progression are species-specific. Whilst in *T. b. gambiense* infection the time between each disease stage averages 3 years and symptoms are often intermittent, *T. b. rhodesiense* infection is rapid, causing an acute disease with little separation between stages and culminating in death within a few months (Malvy & Chappuis 2011). Furthermore, whilst in *T. b. gambiense* infections a

chancre rarely develops at the tsetse bite site, this happens in *T. b. rhodesiense*. The chancre is characterised by general local inflammation symptoms, such as redness, swelling, and pain, but no suppuration (Malvy & Chappuis 2011).

AAT is caused by *T. congolense*, *T. vivax*, and *T. brucei* spp. Whilst in wild animals these parasites cause only a mild infection, domestic cattle, camels, and horses tend to be severely affected after an incubation period varying from 4 days to approximately 8 weeks. Common symptoms range from fever, hair loss, stupor, oedema, anaemia, eye discharge, keratitis, weight loss, abortion and eventually death. Nagana owes its name to the Zulu word “N’gana”, meaning “useless”, because it is a progressive, wasting disease, and causing the animals to become less and less active. For instance, in high-risk areas, animal productivity can be reduced by 38 % (Swallow 1999). Additionally, milk and animal off-take is estimated to be reduced by 8-12 % and 4-10 %, respectively (Swallow 1999). Furthermore, bones lose density, and neurological signs have also been reported. Trypanosomes can also cause immunosuppression, making the host more susceptible to co-infections, such as secondary bacterial infections (Steverding 2008). *T. congolense* is the most virulent aetiological agent of AAT in West Africa, relating to higher levels of anaemia (Dayo et al. 2010).

AAT can be acute or chronic, but the latter is more common in enzootic areas, where prevalence is higher. External factors such as malnourishment, pregnancy-related stress, lactation, excessive work, co-infections, and water deprivation may worsen the disease, but virulence is dependent on both the parasite and host species (Steverding 2008). For instance, trypanosome-susceptible cattle produce lower levels of IgG against trypanosome-specific antigens than tolerant cattle, a phenomenon accentuated during re-infections (Authié et al. 1993).

1.5 Diagnosis

The diagnosis of trypanosomiasis is particularly hard due to non-specific symptoms, low parasitaemia, and abundance of subclinical infections, which act as persistent reservoirs of disease. In most tsetse-infected areas, signs of trypanosomiasis are well recognized, although co-infection with other parasites may mask trypanosomiasis. If available, treatment is given at the first sign of disease, using the

response to therapy as a retrospective diagnosis (Connor 1992). This approach results in poor drug resource management and increased drug resistance.



Figure 3 Sample collection in microhaematocrit capillaries for assessment of blood PCV and examination of wet preparations of the buffy coat. Capillary tubes are filled with animal blood from the ear and spun for diagnosis. Personal photograph (Kenya, 2016).

Antigen variation strongly impairs detection of trypanosomes in the blood by immunological tests and therefore several indirect diagnostic tests have been developed to facilitate fieldwork. Anti-trypanosomal antibodies cannot be used as diagnostic markers since they do not distinguish between current and past infection and cross-reaction between trypanosome species is common. The adapted card agglutination test (CATT) used to assist in the diagnosis of West African human sleeping sickness is also not suitable for AAT (Luckins 1992). The most successful tools are the indirect fluorescent antibody test (IFAT) and the enzyme linked immunosorbent assay (ELISA), always performed in conjunction with another method (Luckins 1992). These require, however, expensive equipment and laboratory expertise.

Recently, several immunochromatographic rapid diagnostic tests have been developed for *T. brucei gambiense*, *T. brucei evansi*, *T. congolense* and *T. vivax*. The first, called HAT Sero-K-SeT, is specific for *T. brucei gambiense* and targets a combination of *T. brucei gambiense*-specific variant surface glycoproteins (Buscher et al. 2013; Büscher et al. 2014); the second, which is called Surra Sero K-SeT, follows the same rationale by targeting the *T. brucei evansi* recombinant VSG RoTat 1.2 (Birhanu et al. 2015). The third, not commercially available yet, is specific for *T. vivax* and targets two invariant surface glycoproteins (TvY486_0045500 and TvY486_0019690); whereas the fourth can successfully diagnose both *T. congolense* and *T. vivax* infections in the same test, is commercially available in Africa, and targets the TcoCB1 antigen, a cathepsin-B protease, and the GM6 antigen a *T. vivax* flagellar-associated protein (**Figure 4**) (Pillay et al. 2013).

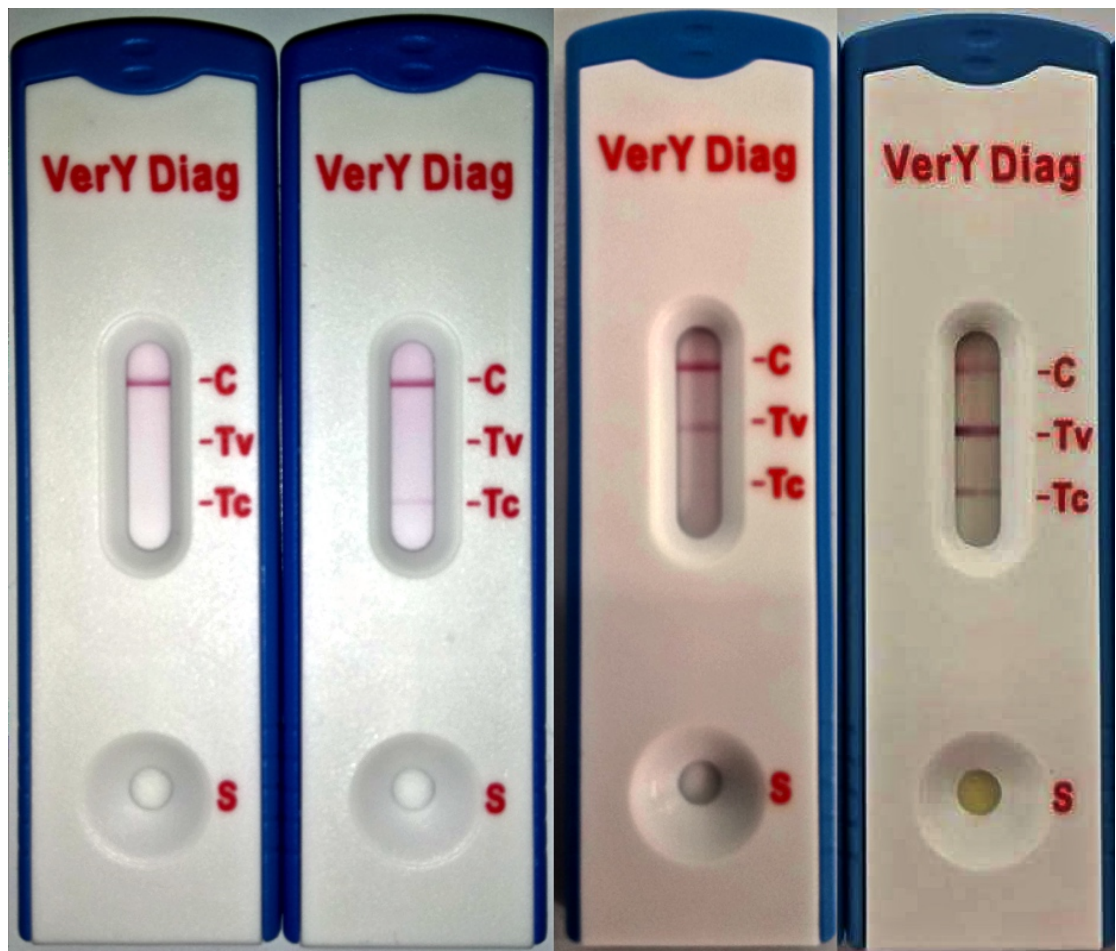


Figure 4 An example of VerY Diag diagnostic test (CEVA, France), which detects *T. vivax* and *T. congolense*, separately or in a multiple infection of the same host (Pillay et al. 2013). Photograph by Alessandra Romero-Ramirez (Liverpool, 2017).

In laboratory settings, species identification can be achieved with restriction fragment length polymorphism (RFLP), isoenzyme electrophoresis, DNA hybridization, and polymerase chain reaction (PCR) assays. Current primers for PCR include *T. congolense* 'savannah' type: IL 0344-0345 (Majiwa et al. 1993), *T. congolense* riverine-forest type TCF 1,2 (Masiga et al. 1992), *T. vivax* VOL 1,2 (Dickin & Gibson 1989), *T. simiae* TSM 1,2 (Masiga et al. 1992), and *T. brucei* s.l. TBR 1,2 (Moser et al. 1989). The sensitivity of the PCR is species-dependent as *T. congolense* achieves the best diagnostic rates to a subspecies level, in contrast with *T. vivax*.

1.6 Treatment

Trypanocidal chemotherapy is usually successful, except in areas of 'savannah' tsetse, due to their high vectorial capacity and higher virulence of carried strains. In these areas, drugs need to be complemented with vector control measures (Connor 1992).

Trypanocides have been available for the past 50 years and include both curative and prophylactic drugs. Curative drugs include diminazene derivatives, quinapyramin sulphate and homidium salts, whereas the only used prophylactic drug is isometamidium (Kroubi et al. 2011). These drugs act on different pathways and therefore have different efficacy rates depending on the trypanosome species. Diminazene has the highest therapeutic index and the broader species range since it binds to DNA, preventing replication. However, it tends to accumulate in tissues, causing a problem for meat consumption. Quinapyramin sulphate, although efficient against *T. congolense*, is not recommended due to cross-resistance mechanisms with other trypanocides (Haroun et al. 2003). Homidium salts are more efficient against *T. vivax* than *T. congolense* and *T. brucei*, but are commonly used as a prophylactic agent.

However, in recent years resistance to trypanocides has been a major problem, particularly in *T. congolense*, due to their prolonged use, lack of dosage control, and high genetic diversity of trypanosomes (Melaku & Birasa 2013). Resistance may be genetically inherited, although is mostly acquired due to drug exposure, cross-resistance, or mutagenesis. Prophylactic drugs are more prone to resistance because of their longer half-life, whereas curative drugs tend to be rapidly eliminated. The predominantly used drugs, isometamidium, homidium and diminazene, are more affected by resistance, which has already been reported in 17 countries of Central Africa (Delespaulx et al. 2008).

The mechanisms of drug resistance in African trypanosomes are not completely understood, but several studies suggest alterations of adenosine transporters involved in the uptake of melaminophenyl arsenicals and diamidines (Barrett & Fairlamb 1999). For example, the loss of *TbAT1* gene expression results in resistance to melarsoprol and diamidines in *T. brucei evansi* and *T. b. brucei*, as this transporter mediates influx of these drugs (Stewart et al. 2010). Mitochondrial electrical potential modulation has also been linked to drug resistance, particularly of

isometamidium (Wilkes et al. 1997). The prolonged use of isometamidium has also been shown to increase resistance to both isometamidium and diminazene of *T. congolense* populations (Peregrine et al. 1997).

1.7 Impact

Efforts to reduce human sleeping sickness have been vastly successful, following a large epidemic in 1998, where 300–500,000 cases were reported. Now, the target of less than 2,000 new cases per year is on track to be reached by 2020 (Franco et al. 2017). Nonetheless, AAT continues to challenge the economic and social development of Central Africa, being a major contributor to poverty. AAT not only has an ecological impact of high cattle mortality rates, but also directly affects Central African economies and demographic distribution, creating persistent, adverse effects on productivity and costing up to US\$4.5 billion per year (HaileMeskel 2016). *T. vivax* has been reported in 11 countries of South America (Venezuela (Clarkson et al. 1971; Garcia et al. 2005), Colombia (Wells et al. 1982; Dirie et al. 1993), Brazil (Shaw & Lainson 1972), Bolivia (Silva et al. 1998), Ecuador (Ortega-Montalvo et al. 2014), Peru (Quispe et al. 2003), Argentina (Monzon et al. 2018), French Guiana, Suriname, Guyana, Paraguay [reviewed by Stephen (1986)], and in five countries of Central America (Costa Rica (Oliveira et al. 2009), Guadeloupe, Martinique, Panama, El Salvador [reviewed by Stephen (1986)]). In Brazil alone, 270M animals are at risk of infection, which usually causes severe epidemics with high mortality rates (up to 40 %) (Silva et al. 1996; Batista et al. 2007; Carvalho et al. 2008; Batista et al. 2009; Batista et al. 2012; de Souza Pimentel et al. 2012; Cadioli et al. 2012; Fávero et al. 2016; Bastos et al. 2017).

1.8 Host-parasite interactions

As obligatory extracellular parasites, African trypanosome bloodstream forms are in constant contact with host tissue fluids where they are challenged by various immunological surveillance mechanisms. Trypanosome survival strategies must rely on complex host cell-parasite interactions to neutralize immunological attacks and ensure parasite survival and longevity (Namangala 2011). Selection pressures have driven trypanosomes to evolve elaborate mechanisms of immune evasion. Combined, they allow the parasite to successfully establish infection and

transmission, while also increasing susceptibility of the host to secondary infections, strongly contributing to the progressive profile of the disease and decreasing responsiveness to vaccination (Namangala, Baetselier, et al. 2000; Shi et al. 2003; Dagenais et al. 2009).

The most remarkable immune evasion mechanism in trypanosomes is antigen variation of VSGs. However, there are other mechanisms, which include generalized immunosuppression targeting macrophages (Namangala, Baetselier, et al. 2000; Gómez-Rodríguez et al. 2009) and T-cells (Jayawardena et al. 1978); induction of complement activation and antigen-presenting cell deficiencies (Namangala, Brys, et al. 2000; Dagenais et al. 2009), suppression of lymphocyte proliferative responses (Sileghem et al. 1994; Radwanska et al. 2008), modulation of T-cell and B-cell activity (Schleifer et al. 1993; Radwanska et al. 2008), and acquisition of resistance to human trypanolytic activity by *T. brucei gambiense* and *T. brucei rhodesiense* (De Greef & Hamers 1994; Van Xong et al. 1998; Uzureau et al. 2013; Capewell et al. 2013). VSGs also play a role in some of these mechanisms. For example, VSGs prevent trypanosome cell complement-mediated lysis (Vincendeau & Bouteille 2006). They also induce continuous cytokine production (particularly TNF- α) and autoantibody synthesis (Tachado & Schofield 1994; Okomo-Assoumou et al. 1995).

Host-parasite interactions are mediated at the surface of the cell, the point of contact between the parasite and the extracellular environment. The surface of African trypanosomes is mostly composed of a glycoprotein monolayer of proteins, which varies with the life stage. In the procyclic form, they express procyclin; in the metacyclic stage, they are coated with mVSG (Pedram & Donelson 1999); and in the bloodstream form, African trypanosomes express bloodstream form VSGs.

However, buried within these coats are multiple surface proteins, the majority of which belong to multi-copy gene families (**Table 2**). In 2013, Jackson et al. presented a three-way comparison of the surface architecture of *T. brucei*, *T. congolense* and *T. vivax* with the aim to identify species-specific genes or gene families that could be linked to the phenotypic variation observed in AAT, be it virulence, pathology, host range, or transmission mode (Jackson et al. 2013). Surface genes families were identified and characterised based on a combination of sequence similarity searches, clustering analyses and phylogenetic estimation and reconciliation (Jackson et al. 2013).

The 'Cell-surface Phylome' revealed that *T. brucei*, *T. congolense* and *T. vivax* share 34 non-VSG gene families encoding surface proteins, which include multiple membrane transporters, invariant surface glycoproteins, major surface protease, trans-sialidase, protein kinases, adenylate cyclase, lipase, cysteine peptidases, few isomerases, and several hypothetical proteins (Jackson et al. 2013). *T. brucei* and *T. congolense* share three additional gene families, which comprise two families of transferrin receptors (Fam14 and 15) and a putative secreted protein (Fam10), reflective of the shorter genetic distance between these species and their procyclic life stage. Of the three species, *T. vivax* has the most complex surface phylome, containing 23 *T. vivax*-specific gene families, of which 19 are non-VSG. These genes are mostly putative secreted or putative membrane proteins, but also include the known mucin-associated surface protein (MASP, Fam35).

This study concluded that the essential features of the African trypanosomes surface are conserved between species, and thus were established in the common ancestor. However, there are prominent gene families that have expanded rapidly in each organism and a set of genes that reflect recent species-specific adaptations (Jackson et al. 2013).

Table 2 The surface of African trypanosomes according to the Cell Surface Phylome (Jackson et al. 2013).

	Single-copy genes	Double-copy genes	Triple-copy genes	Multi-copy gene families	Description
<i>T. brucei</i> only	101	6	4	10	Fam0-9
<i>T. congolense</i> only	153	44	10	6	Fam16-22
<i>T. vivax</i> only	214	31	19	23	Fam23-45
<i>T. brucei</i> + <i>T. congolense</i>	N/A	N/A	N/A	5	Fam10-15
All	N/A	N/A	N/A	34	Fam46-81

1.9 Antigenic variation

Antigenic variation is a mechanism of immune evasion used by several pathogens, such as African trypanosomes, *Babesia*, *Plasmodium*, *Giardia*, *Neisseria gonorrhoeae*, *Borrelia hermsii*, *Anaplasma*, and *Pneumocystis*, consisting of the sequential substitution of cell surface antigens, which coat the whole surface of the pathogen.

In African trypanosomes, antigenic variation was first documented in the early 20th century, through the observation of the survival of specific subsets of trypanosomes in otherwise lethal sera [reviewed by Soltys (1963)]. This led to the discovery of 'variable antigen types' (VATs), which elucidated the trypanosome survival strategy in the mammal host. The sequential expansion of particular VATs resulted in 'waves of parasitaemia', cyclical fever and long-lasting infection (Barry & McCulloch 2001). *Trypanosoma brucei* is the model organism for antigenic variation (Deitsch et al. 2009).

VSGs are highly antigenic and induce a strong B-cell dependent immune response. However, infection persists because clearance of VSG-bound antibodies results in the proliferation of parasite variants that express a different VSG. This process results in characteristic 'waves of parasitaemia', where parasitaemia peaks are followed by periods of undetectable parasite levels (**Figure 5a**). If untreated, the cycle persists until exhaustion of the host immune system or of the antigenically distinct VSG variants.

VSGs are expressed from dedicated cassettes at the telomeres of megabase chromosomes called VSG expression sites (ES) (Berriman et al. 2002; Becker et al. 2004; Hertz-Fowler et al. 2008) (see section 1.10.3). To avoid parasite clearance by anti-VSG antibodies, the parasite has the ability to switch the monoexpressed VSG to one of the hundreds of VSGs in the genome. VSG switching occurs following transposition of another VSG to the ES (a process known as gene conversion), or through activation of a different ES (see section 1.10.4). Both processes result in VSG replacement in the parasite surface, independently of the host immune pressure. This mechanism successfully impairs antibody-mediated lysis (Balber et al. 1979) and promotes clearance of VSG-bound antibodies (Engstler et al. 2007).

Whilst traditionally each peak of parasitaemia has been associated with clonal expansion of a single variant, Hall et al. (2013) showed that, in *T. brucei*, they comprise a minimum of 15 variants, indicating that the waves of parasitaemia result from expansion of antigenically-distinct populations. Furthermore, Mugnier et al. (2015) showed that, in *T. brucei*, the ‘waves of parasitaemia’ are restricted to early infection and not as radical as depicted in **Figure 5a**. The number of VSGs being expressed by the population during infection increases disconcertedly; therefore, although defined waves of parasitaemia are observed at the individual variant level, the total parasitaemia never drops abruptly (**Figure 5b**). In late infections, this effect is accentuated so the total parasitaemia becomes almost constantly high.

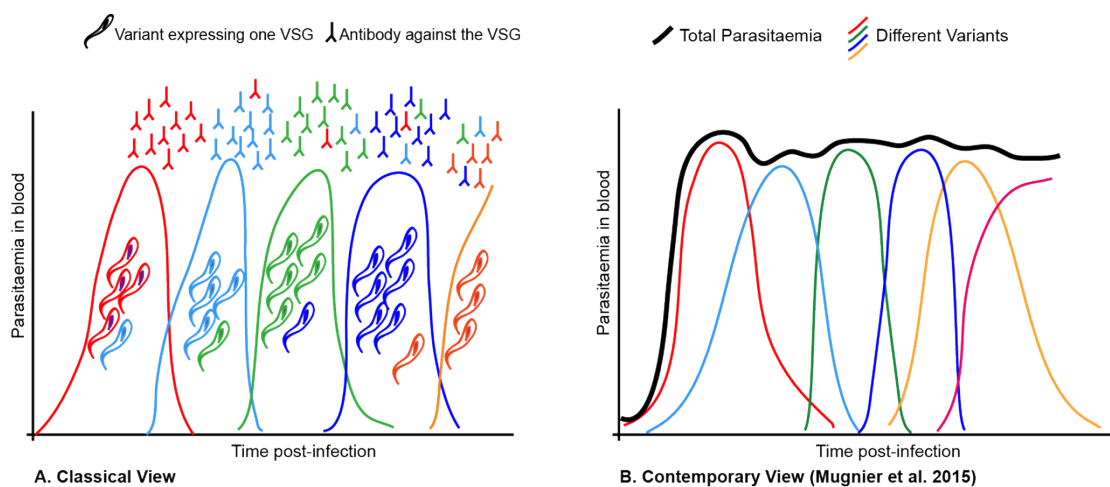


Figure 5 Host-Parasite interactions: ‘waves of parasitaemia’ in a trypanosome infection. A. Classical View: sequential growth and decay of five trypanosome variants expressing different VSGs. Clonal populations of parasites rapidly proliferate, triggering a strong antibody response against the expressed VSG and resulting in population clearing. Meanwhile, minority parasite populations expressing a different VSG can proliferate until a second antibody response is triggered. Theoretically, the cycle can continue until exhaustion of the host immune system. B. Contemporary View: Waves of total parasitaemia are due to multiple variants (Hall et al. 2013) and are less obvious because the emergence of alternative variants is disconcerted (Mugnier et al. 2015).

1.10 Variant Surface Glycoproteins

1.10.1 VSG structure

The VSG surface coat has the main function of protecting the trypanosome population against the host immune response. However, as the mammal hosts elicit

innate (non-specific) and adaptive (specific) immune responses, VSGs are under two opposite pressures: structural conservation so that the invariant surface molecules remain protected from the immune effectors, and epitope variation so that each VSG coat is sufficiently distinct to undergo antigenic variation and elicit a new antibody response. To achieve these demands, the VSG tertiary structure is conserved (Carrington et al. 1991; Blum et al. 1993) and they organise in a densely packed monolayer that protects against complement mediated lysis (**Figure 6**). Each glycoprotein is attached to the cell membrane by a glycosylphosphatidylinositol (GPI) anchor. Buried within the VSG coat, and protected from the immune system, are invariant molecules.

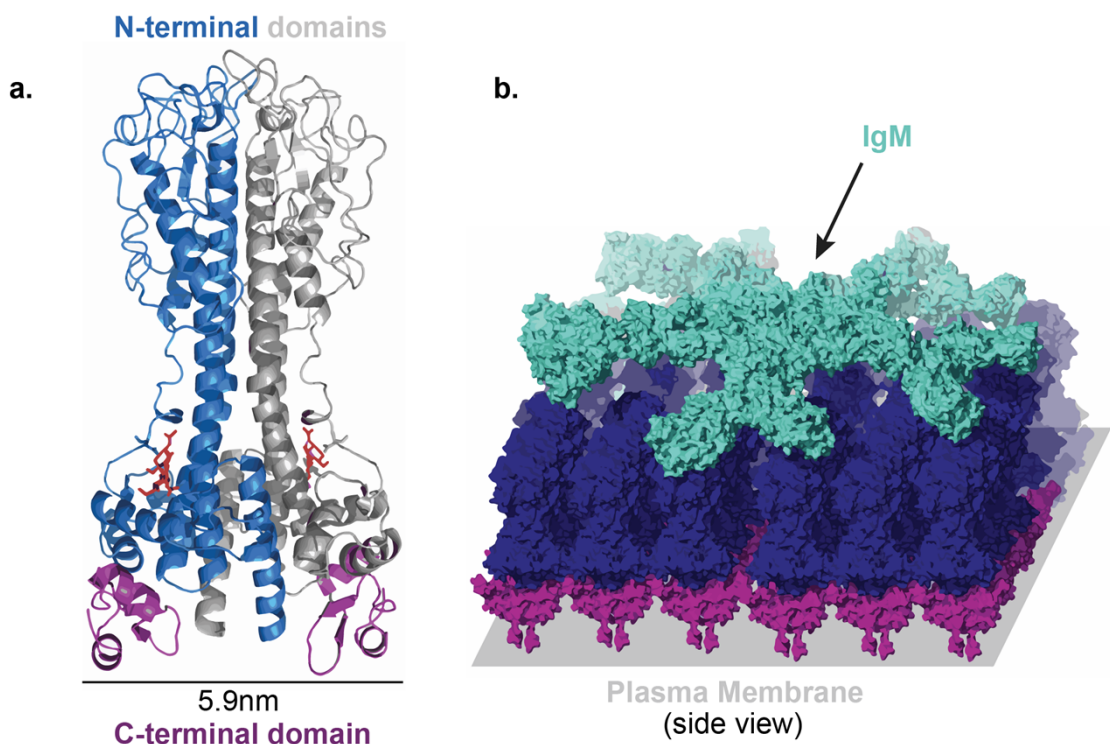


Figure 6 The VSG structure. a) The 3D model of VSG221 dimer from *T. brucei*. N-terminal domains are shown in blue and grey, representing each monomer. C-terminal domains are shown in purple. Adapted from Schwede et al. (2015). b) **Hypothetical model of antibody (IgM) binding to VSG.** The model reflects the dense packing of the plasma membrane. VSG N-terminal domains are shown in blue and the C-terminal domains in purple. Adapted from Mugnier et al. (2016).

VSG have a long highly variable N-terminus, with two antiparallel α -helices separated by a turn, and a shorter conserved C-terminus that anchors to the plasma membrane through a GPI anchor (**Figure 6**). VSG are often N-glycosylated, which

may impact the structure and accessibility of the VSG epitopes or adjacent molecules (Mehlert et al. 2002), and contribute to dimensional conservation between variants (Blum et al. 1993). After translation, the VSG is coupled to a 30-40 amino acid signal peptide at the N-terminal domain, whose cleavage directs the protein for further post-translational modification, including GPI anchor addition and N-linked glycosylation (Ferguson et al. 1986; Mehlert et al. 2002). After that, the VSG is directed to the flagellar pocket to reach the plasma membrane. VSG molecules are indeed subject to rigorous quality control steps. A series of mutations in two *T. brucei* VSG drastically reduced the expression levels of the proteins, even though the GPI anchor was maintained (Wang et al. 2003). An efficient display on the parasite surface seems to require all conserved structural motifs, such as the cysteines, tryptophans, and the N-glycosylation sites.

Most VSGs are clustered in subtelomeric arrays in the megabase chromosomes and these genes comprise approximately 20 % of the whole genome (Berriman et al. 2005). Although African trypanosomes have mostly diploid genomes (Hope et al. 1999; Peacock, Ferris, et al. 2014), the subtelomeres are hemizygous, making individual VSGs haploid (Callejas et al. 2006; Marcello & Barry 2007a).

The VSG repertoire is the major source of genetic diversity in the parasite, although its size varies by species and strain (Donelson 2003; Cross et al. 2014). The physical sequence properties of each species VSGs are distinct. *T. brucei* VSGs are normally the longest (mean length = 498 ± 29 amino acids), whereas *T. congolense* and *T. vivax* VSG are shorter (mean length = 388 ± 30 and 394 ± 95 respectively) (Jackson et al. 2012). These length differences are due to *T. brucei* VSG having longer hypervariable regions at the N-terminal domain (NTD) and a much longer C-terminal domain (CTD). In terms of physical properties, *T. brucei* and *T. congolense* VSG have similar GC contents (pGC = 0.488 ± 0.016 and 0.481 ± 0.032 , respectively). *T. vivax* VSGs have a higher pGC of 0.599 ± 0.019 , although the genome-wide GC content is also higher. Yet, *T. vivax* VSG have a higher codon bias towards hydrophobic and less aromatic amino acids, which might contribute to the disparity in VSG GC content (Jackson et al. 2012). The number of pseudogenic, frameshifted or degenerate VSGs is also distinct between species. While less than 20 % of *T. brucei* spp. are full-length VSGs, this number rises to approximately 75 % in *T. congolense* and *T. vivax* (Jackson et al. 2012). In all three species, VSGs are mostly arranged in subtelomeric tandem arrays and in the mini-chromosomes. However, in *T. congolense*, the VSG arrays often contain non-VSG multi-copy

genes also, such as the transferrin receptors (Fam14/Fam15) and invariant surface glycoproteins.

1.10.2 VSG diversity

VSGs are highly diverse, particularly in their NTDs due to their exposure to the host immune pressures. The primary structure of VSGs varies considerably between African trypanosome species. Nonetheless, the pattern of conserved cysteine, glycine and tryptophan residues reported for *T. brucei* VSG are also observed in *T. congolense* and *T. vivax* (Carrington et al. 1991; Blum et al. 1993; Berriman et al. 2005; Jackson et al. 2013). Two subgroups have been identified, a-VSG and b-VSG. These groups are characterised by specific amino acid motifs, such as the 'GRIDE' motif of a-VSG (Salmon et al. 1997), of which the transferrin receptors (TFR) are members (Jackson et al. 2013), and the central CxC motif of b-VSG (Blum et al. 1993; Jackson et al. 2013).

In 2012, Jackson et al. provided a detailed comparison of the global VSG repertoires of *T. brucei*, *T. congolense*, and *T. vivax* with the aim to understand how antigenic diversity has evolved in the trypanosome genomes (Jackson et al. 2012). This study revealed that *T. brucei* a-VSGs are more closely related to a-VSG-like subfamilies in both *T. congolense* (TFR) and *T. vivax* (Fam23) than to *T. brucei* b-VSGs. Similarly, *T. brucei* b-VSGs are more similar to b-VSG-like subfamilies in *T. congolense* (Fam13 and Fam16) and *T. vivax* (Fam24) (**Table 3**). These results indicate that VSG lineages precede speciation events, suggesting that the genome of their last common ancestor contained both a-type and b-type VSG lineages that were inherited by the contemporary species (Jackson et al. 2012).

Table 3 Summary of VSG-like genes in African trypanosomes (Jackson et al. 2012).

	<i>T. brucei</i>	<i>T. congolense</i>	<i>T. vivax</i>
a-type	a-VSG	-	Fam23
	Fam15 (Transferrin Receptor)	Fam15	-
	Fam14 (Procyclin-associated genes)	Fam14	-
b-type	b-VSG	-	Fam24
	Fam9 (VSG-related genes)	Fam16	-
	Expression site-associated gene 2 (ESAG2)	Fam13	-
	Fam1 (VSG-related genes)	-	-
other	-	-	Fam25
			Fam26

Within a- and b-VSG, there is only 10-15 % peptide conservation at the N-terminus, comprising a vast reservoir of antigens. However, there is also little homology between VSG of the same type but of different species (Strickler et al. 1987). Though a-VSG account for 50 % of the *T. brucei* VSG repertoire and for more than 500 genes in *T. vivax*, a-type variant antigens are absent from the *T. congolense* genome (Helm et al. 2009; Jackson et al. 2013). In *T. congolense*, the a-type VSG family is represented exclusively by the non-variant transferrin receptors (Fam14/15) (**Table 3**).

T. brucei b-type VSG are a diverse family, which share a common C-terminal domain. By contrast, *T. congolense* b-type VSG cluster within two structurally different families (Fam13 and Fam16), likely to have originated in the African trypanosome ancestor because their closest relative are VSG-like genes in *T. brucei* (ESAG2 and VR9, respectively), rather than each other (**Table 3**). For this reason, *T. brucei* b-type VSG have been suggested to have passed through a bottleneck evolution from a single ancestral lineage arising after speciation (Jackson et al. 2012).

The *T. congolense* VSG families are further split into 15-20 phylotypes with non-homologous CTDs, all non-homologous to *T. brucei*. Although the CTD of *T. brucei* is conserved across a- and b-VSG, so it can work as recombination anchor point for VSG movement between the subtelomeres and the expression sites, *T. congolense* b-VSG have at least 15 types of CTD, and for *T. vivax* an obvious CTD has yet to be

identified. Relative to *T. brucei*, *T. vivax* presents the greatest VSG structural diversity, consistent with its early separation from the African trypanosomes phylogenetic branch. Its VSG repertoire is divided into four subfamilies (Fam23–26), of which two are homologous to a-VSG and b-VSG (Fam23, Fam24) and two are species-specific (Fam25, Fam26). It is not clear whether these two *T. vivax*-specific families are functional VSG and whether they are derived from ancestral lineages that were lost in *T. congolense* and *T. brucei* or have originated post-speciation in *T. vivax*. Sequence variation within these families is very low, although diversity between lineages is remarkably high.

The cladistic structure of *T. congolense* and *T. vivax* phylogenies suggests wider distances between clades and the conservation of different ancestor lineages information within basal nodes (**Figure 7**). The *T. brucei* VSG phylogeny contrasts with the latter, showing genome-specific long branches with narrower distributions and rare basal internodes (**Figure 7**). This dispersed structure suggests high levels of recombination, which is not so likely in *T. congolense*, mostly due to the conserved CTDs, but also due to the evolutionary barrier of a wider distribution of sequence variation. *T. vivax*, on its side, has the lowest rate of VSG recombination which might be explained by the same reasons as *T. congolense* in combination with differences in reproduction modes: whereas *T. brucei* and *T. congolense* seem to undergo sexual recombination, *T. vivax* may exclusively rely on clonal reproduction (Tait & Turner 1990; Tait et al. 2007; Peacock et al. 2009; Morrison, Tweedie, et al. 2009; Duffy et al. 2009; Tihon et al. 2017).

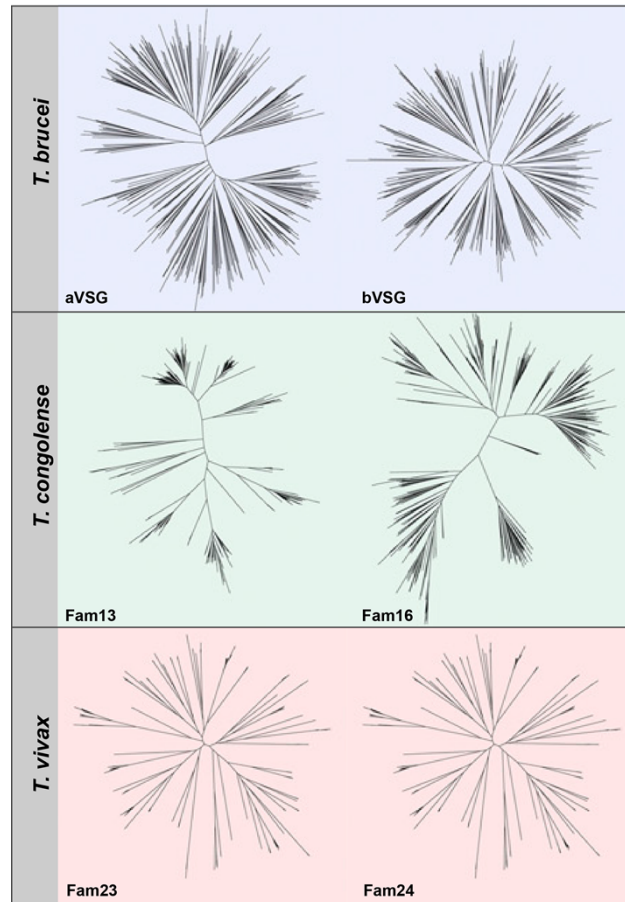


Figure 7 Phylogenetic tree topologies for VSG-like subfamilies in *T. brucei*, *T. congolense* and *T. vivax*. Differences in topologies derive from variation in the ratio of internal to terminal branches ('treeness'). Treeness is highest for *T. vivax* (0.681 and 0.763), and lowest for *T. brucei* (0.282 and 0.275). All trees are drawn to the same scale. Adapted from Jackson et al. (2012).

While Fam13 and Fam16 in *T. congolense*, and Fam23 and 26 in *T. vivax* contain VSGs with variant antigen functions (Strickler et al. 1987; Rausch et al. 1994; Eshita et al. 1992; Helm et al. 2009; Jackson et al. 2013; Jackson et al. 2015), it remains unknown whether all VSG homologues encode variant antigens. In *T. brucei*, the VSG repertoire contains closely related genes, such as ESAG2, VR and other VSG-like genes, which have evolved non-variant roles. Analogous families in *T. congolense* and *T. vivax* have not been identified to date.

The system of antigenic variation has evolved as an adaptive mechanism to create multiple phenotypes from a clonal population and thus evade immune responses (Barry & McCulloch 2001). Therefore, the mechanisms of antigenic variation must

include a fine-tuned machinery to allow a rapid change of the surface antigen from a gene repertoire large enough that prevents antigenic exhaustion. An example of how this can be achieved is the existence of a vast silent gene 'archive' (Deitsch et al. 2009), that works as a library of antigenic diversity. The potential for variation can be increased by recombinatorial mechanisms that allow the combination of segments of distinct silent genes to create novel, mosaic, and antigenically distinct genes during antigenic switching (see section 1.7.3). Depending on the efficiency and frequency of these mechanisms, they have the ability to transform a small archive into a much larger antigenic pool (Zhuang et al. 2007).

1.10.3 VSG expression

In the *T. brucei* metacyclic stage, VSG are expressed from dedicated cassettes located at the ends of the longer chromosomes, called the metacyclic expression sites (MES). The MES comprises a long transcription promoter of 426 bp, a pyrimidine-rich region, the VSG, and a short 70 bp repeat, which facilitates gene conversion to the expression site after infecting of the host (Graham & Barry 1995). The mVSG expressed by a parasite population are heterogeneous, but limited to up to 27 antigen types (Turner et al. 1988). Nonetheless, mVSG expression is essential during the first five days of infection. This repertoire has been shown to change gradually over time and cyclical transmissions, and varies between strains (Barry et al. 1983). Like in *T. brucei*, the *T. congolense* mVSG population is heterogeneous and its repertoire also thought to be limited to a specific set of antigen types, which remain expressed up to 9 days post-infection in the mammal host (Crowe et al. 1983). The relationship between these antigen types and the *T. congolense* VSG phylotypes remains unclear due to the previous inability to dissect the antigen types to the gene level.

Within seven days of infection, as the metacyclic parasites differentiate into bloodstream forms, the mVSG is replaced by a bloodstream VSG. In the bloodstream life stages, VSG must be produced in very high quantities to cover the full surface of the parasite. In fact, 10 % of the total soluble protein of *T. brucei* bloodstream forms is VSG (Cross 1990). To produce such quantities of protein, a large number of VSG mRNA transcripts are required. VSG and its associated genes are transcribed from a specialised telomeric expression site (ES) called the bloodstream ES (BES) by RNA polymerase I (Pol I), representing the sole exception of Pol I transcribing mRNA in eukaryotes (Günzl et al. 2003). The *T. brucei*

bloodstream-form VSG ES are specific telomere-proximal transcription units in megabase and intermediate chromosomes (Berriman et al. 2002; Hertz-Fowler et al. 2008), which contain the VSG and a collection of ESAGs that are co-transcribed with the VSG. Thirteen families of ESAGs have been identified, all in association with VSG in the ES. Although some appear to be constitutively present, such as ESAG6 and ESAG7, ES usually harbour only 5 to 10 ESAGs and pseudo-ESAGs, which include degenerate, frameshifted, and truncated genes (Donelson 2003; Berriman et al. 2005).

The *T. brucei* ESAGs derived uniquely in *T. brucei* from conserved multi-copy gene families found in the core chromosomes or subtelomeres. They have been independently recruited to the expression site, where they are subject to higher sequence recombination, possibly from telomeric exchange, resulting in concerted evolution between these distinct genes (Hertz-Fowler et al. 2008; Jackson et al. 2013). Up to 90 % of the VSG are coupled to an upstream regulatory 70 bp repeat region known to facilitate VSG transposition between the subtelomeres and the expression sites. In *T. congolense*, regions homologous to ESAGs and expression sites have not yet been found. Extensive telomeric and subtelomeric cloning in *T. congolense* has become an area of great interest with the goal of understanding VSG expression, switching and recombination. Phylogenetic analysis of ESAGs suggests that they are a *T. brucei*-specific innovation (Jackson et al. 2013), but it is plausible that *T. congolense* recruited the same or other genes to the expression sites to perform ESAG functions.

In *T. brucei*, only one VSG is actively transcribed at any given time. VSG monoallelic expression is advantageous because it results in a homogeneous parasite surface coat, which exposes a single antigen at a time to immune system, precluding antigen repertoire exhaustion [reviewed by Horn (2014)]. The VSG is expressed from an 'active ES' [reviewed by Horn (2014)].

Although 'double-expressers', or parasites expressing two VSG at a time, can be created in the experimental settings, this is unprecedented in the wild (Muñoz-Jordán et al. 1996). ES activation is rigorously controlled and various experiments have shown that in chemically induced activation of two ES, the parasites become unstable, but show highly heritable patterns of dynamic ES switching (Chaves et al. 1999). The active ES co-localises with the extranucleolar, Pol I-containing expression site body (ESB) (Navarro & Gull 2001) and their association is usually

inherited, suggesting that monoallelic expression is inherited and epigenetically controlled. Epigenetic mechanisms identified to date include chromatin remodelling [e.g the reduction of TblSWI (Hughes et al. 2007; Stanne et al. 2015) and histone H1 (Pena et al. 2014)] and nuclear location, as inactive ES, all located at the nuclear periphery, where interaction with NUP1 ensures ES silencing (DuBois et al. 2012). Furthermore, nucleosome depletion and chromatin remodelling (particularly histone H1) play an important role in Pol I control (Figueiredo & Cross 2010; Pena et al. 2014), which in turn might be actively involved in monoallelic exclusion control, as it has been recently shown that Pol I inhibition causes fragmentation of the nucleolus and loss of ESB (Kerry et al. 2017).

1.10.4 VSG switching and recycling

Despite the mechanism of monoallelic expression, it must be flexible enough to allow for rapid and frequent switching. VSG recycling occurs at the flagellar pocket, the only site of exocytosis and endocytosis. This is a rapid process lasting for about 12 min that allows the replacement of VSG molecules in the surface coat, functioning as a cleaning mechanism and facilitating trypanosomal survival in low titres of anti-VSG antibodies.

VSG switching is the autonomous and spontaneous process of changing the expressed VSG coat independently of the immune system (Doyle et al. 1980). In the wild, VSG switching in *T. brucei* occurs at a rate of 10^{-2} to 10^{-3} switches per cell per population, although in laboratory-adapted strains the rate drops to about 10^{-6} (Donelson 2003). VSG switching can occur through two main mechanisms: transcriptional and recombinatorial switching. The former is achieved through the transfer of monoallelic expression to another ES, without the need for genetic rearrangement. The latter involves the copying or transposition of a VSG into the ES and can occur through gene conversion, segmental gene conversion (SGC) and telomere exchange (**Figure 8**).

Transcriptional switching consists of the silencing of the active ES and subsequent activation of a different ES (Majiwa et al. 1982; Young et al. 1982; Young et al. 1983) (**Figure 8A**). It has been proposed that this mechanism occurs through the transference of the ESB to a different ES (Navarro & Gull 2001), although it may be more complex, as the activation of the new ES and the repression of the previous ES are independent processes (Figueiredo et al. 2008). Regardless of its players, *in*

situ VSG switching is considered of minor importance for natural infections and pleomorphic cell lines (Robinson et al. 1999). In fact, the main mechanism of VSG switching is (duplicative) gene conversion, where VSG are duplicated or transposed from the subtelomeric arrays to the active ES by homologous recombination (**Figure 8B**). This process is usually triggered by damage to the ES, often by double-stranded DNA breaks (Morrison, Marcello, et al. 2009; Boothroyd et al. 2009; Glover et al. 2013), resulting in the deletion of the original DNA sequence and its replacement with the template strand.

For homologous recombination to occur, two DNA strands must have at least two regions similar enough in sequence to allow for strand annealing and binding of the newly-synthesised sequence to the ES sequence being repaired. Thus, the regions involved in VSG conversion are the 70 bp repeat upstream the VSG and its C-terminal domain or 3' UTR, which work as 'anchor points' for the VSG switching (Bernards et al. 1981; Liu et al. 1983; Liu et al. 1985; Timmers et al. 1987). Whilst duplicative gene conversion requires exchange of intact VSG, the vast majority of *T. brucei* VSGs are frameshifted, interspaced by stop codons, or degenerate. These VSGs can be repaired or modified into functional VSG by SGC (**Figure 8C**), which can occur through mosaicism and 3' donation (Barbet & Kamper 1993). Mosaic VSG are formed through the combination of segments from multiple donors and can contribute not only to a greater use of the VSG archive, but also to an exponential increase in antigenic variation potential (Hall et al. 2013). Expression of mosaic VSG may be evolutionary favoured due to the pressure for diversity generation imposed by the host adaptive immune system, as they can present a massive benefit in infection chronicity (Barbet & Kamper 1993).

The exact mechanisms of SGC are unclear, although likely to be based on sequence homology as mosaic VSGs and their donors often share sequence similarities. As the flanking regions that allow classical homologous recombination are lacking (Barnes & McCulloch 2007) and the segments can be as small as 14 base pairs (Kamper & Barbet 1992), it was proposed that mosaic VSG were formed by series of recombinatorial events, also known as 'progressive mosaicism' (Pays, Houard, et al. 1985; Barry et al. 2005; Marcello & Barry 2007a). This hypothesis has been corroborated by recent experiments. Hall et al. (2013) confirmed that mosaicism plays a large role in antigenic variation in *T. brucei*, as sequence identity does not constrain gene conversion; and provided evidence that antigenic variation is built continuously by many variants, rather than in a tightly regulated, intermittent

manner. Moreover, it was shown for the first time that SGC does repair pseudogenic donors, and creates antigenically distinct variants, contributing to great richness of VSG expression (Hall et al. 2013). Through rapid accumulation of mosaicism, expressed VSGs have a new layer of diversity that exponentially increases the potential for antigenic variation.

The last known mechanism of VSG switching is telomeric exchange, through which classical recombination allows the exchange of two chromosome ends, resulting in the substitution of the telomere-proximal VSG and its downstream sequence (Pays, Staerz, et al. 1985) (**Figure 8D**).

It is believed, although not yet shown experimentally, that the switching rate is dependent on a cyclical transmission through tsetse flies, possibly to ensure a labile epigenetic state. Regardless of the mechanism, the rapid process of antigenic switching is always followed by the slower replacement of the VSG coat. The whole replacement of the VSG coat is thought to take two days *in vivo*, a theoretical number that accounts for the rapid gene expression change, the VSG mRNA half life of 4.5 hours, and the VSG half life of 30 hours (Cross et al. 1998). The veracity of this assumption remains to be empirically tested, although, in *T. brucei*, the change between mVSG and bloodstream VSG has been shown to take 2 days, by detection with monoclonal antibodies (Esser & Schoenbecher 1985).

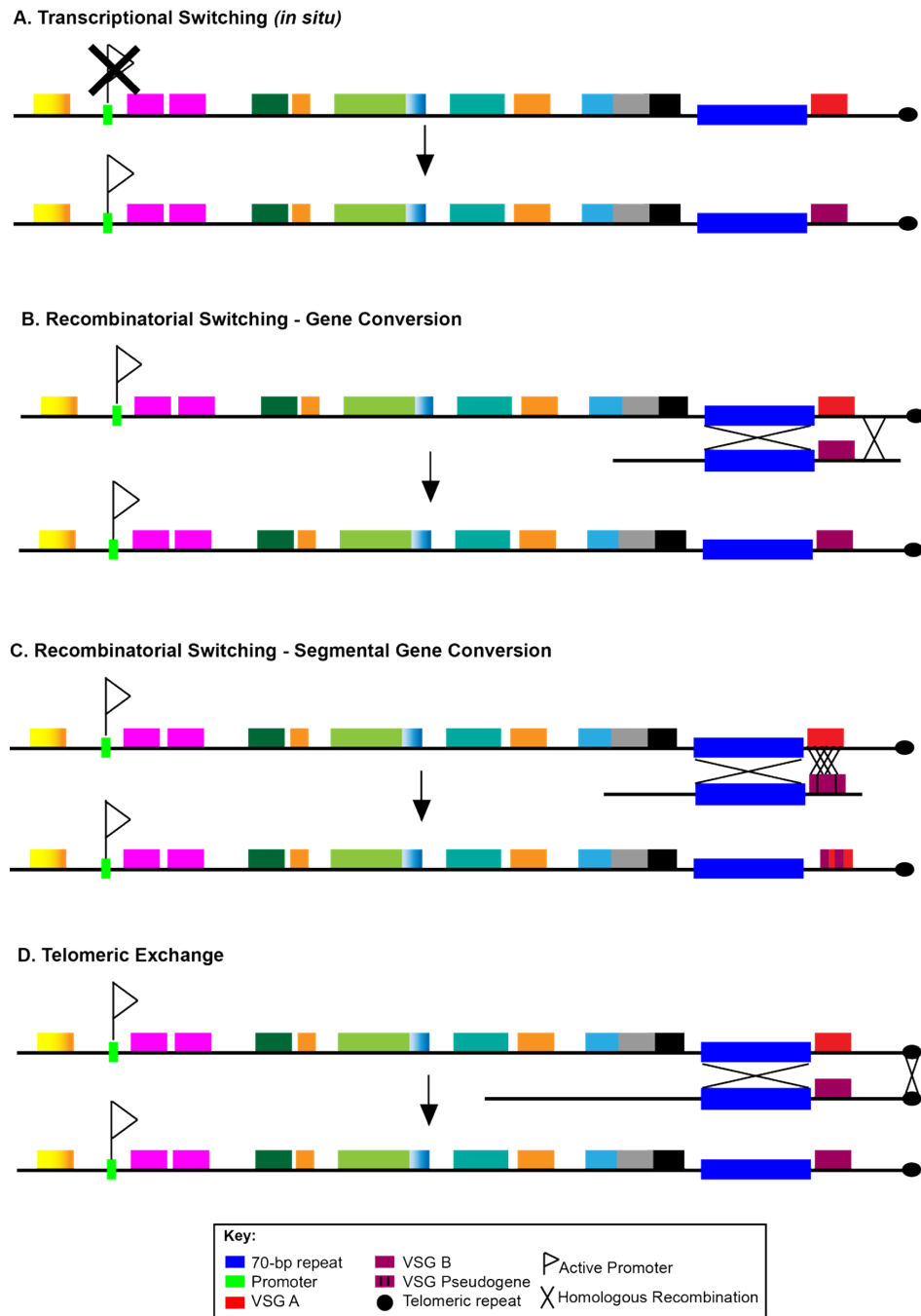


Figure 8 Mechanisms of VSG switching. A. Transcriptional or *in situ* switching consists on the activation of a different expression site (ES), without any gene rearrangement. B. Recombinatorial Switching can be achieved through gene conversion when an intact VSG is transferred to the active ES by homologous recombination usually between the 70 bp repeat regions upstream of the VSG and its C-terminal domain. C. Recombinatorial Switching can also be achieved by segmental gene conversion, which occurs when segments of a silent pseudogenic VSG are transferred to the active expression site, creating a novel, mosaic VSG consisting of segments of the original and the pseudogenic VSG. D. The mechanism of telomeric exchange by homologous recombination can also contribute to VSG switching.

1.10.5 Hierarchic activation of VSGs

VSG activation is partially hierarchical. The order of the expressed VSG in the infrapopulation (i.e. the trypanosome population of a given host at any given time) is somewhat predictable, as specific VSGs tend to be expressed at particular times of infection. This hierarchy, which is observed in other pathogens, such as *Plasmodium* and *Anaplasma marginale* (Bull et al. 2005; Chávez et al. 2012), helps to prevent rapid exposure of the full VSG archive to the immune system (Morrison et al. 2005), but it may also contribute to maintain the infection at a sub-lethal stage, as spontaneous antigen variability could lead to extreme parasitaemia and immune exhaustion of the host (Turner et al. 1995). Ordered VSG activation in independent *T. brucei* infections was first observed by (Gray 1965), but has since been supported by several studies, which showed that similar subsets of antigen types appear in early infections (McNeillage et al. 1969; Van Meirvenne et al. 1975; Miller & Turner 1981; Liu et al. 1985). Longer infections in *T. brucei equiperdum* (one month) (Capbern et al. 1977) and *T. vivax* (several months) (Barry 1986) further confirmed the conserved order of VSG expression, although suggested that the pattern becomes looser as infection progresses.

The mechanisms of expression hierarchy is based on the concept of VSGs having distinct activation probabilities (Pays 1989). The activation probability of a VSG is thought to be partly related to its genomic locus, as several studies have observed a correlation between that and the timing of activation (Young et al. 1983; Liu et al. 1985; Robinson et al. 1999). For instance, in *T. brucei* infections, minichromosomal VSGs arise early in infection, activated mainly by duplicative gene conversion between the active expression site and other telomeres, whereas VSGs from subtelomeric arrays are activated less often and mostly in late infections.

Morrison et al. (2005) tested the degree of predictability in VAT expression and its association with the VSG locus type, showing that the VSG activation probability is directly linked to the complexity of moving it into the expression site. As gene conversion is homology-based, it is possible that the sequence similarity between the VSG being activated (and its flanking regions) and the VSG being silenced contribute to the activation probability (Morrison et al. 2005). As the repetitive and unstable nature of the telomeres favours recombination (Barry et al. 2003), it is sensible that telomeric VSG are more likely to be activated, and therefore appear

early in infection (Liu et al. 1985; Robinson et al. 1999). Accordingly, full VSGs from subtelomeric arrays are second in the order of activation probability as they can be transposed to the expression site in a single gene conversion step based on homologous recombination between the 70 bp repeat and the 3' end of the VSG (Marcello & Barry 2007a). Finally, pseudogenic VSGs have the lowest activation probability as SGC requires multiple steps of repairing recombination; therefore they tend to appear in later infections, playing an important role in infection chronicity rather than acute disease (Pays 1989; Hall et al. 2013). Nonetheless, Hall et al. (2013) have shown that the hierarchical activation is loose, not strict. Although the frequency of mosaic VSGs increases as infection progresses, they can be expressed earlier, particularly due to SGC modifications to early-expressed VSGs.

1.10.6 Role of VSGs in pathology and virulence

Antibodies against VSGs are the key effectors of the host adaptive immune response to trypanosomes. As the surface of bloodstream form trypanosome is coated with a fluid, tightly packed, highly immunogenic VSG monolayer, the immune system recognises the conformational epitopes of the N-terminal domain and elicits an antibody response (Mehrlert et al. 2002; Pinder et al. 1987; Miller et al. 1984a; Miller et al. 1984b; Masterson et al. 1988). Furthermore, VSGs can be secreted from the parasite. Soluble VSGs are not very immunogenic, i.e. do not bind antibodies strongly (Black et al. 1982), but may assist in B-cell activation and/or alternative immune responses, such as complemented-activated cascades (Sendashonga & Black 1982; Black et al. 1982). Frequent recombination and mutagenesis generate vast antibody diversity. Therefore, trypanosomes have evolved mechanisms of generating antigenic diversity to evade it. Simultaneously, trypanosomes induce immunosuppression through the activity of suppressor macrophages and induction of apoptosis in marginal zone B-cells (Radwanska et al. 2008). Besides the extrinsic effect exerted by the immune system in the parasite population, in *T. brucei*, there is an intrinsic population control that prevents abrupt parasite growth and host immune exhaustion. A portion of slender bloodstream forms differentiates into tsetse-infective quiescent stumpy forms, stabilising parasitaemia and amplifying the chance of successful tsetse infection. This developmental mechanism is linked to repression of VSG transcription (Amiguet-Vercher et al. 2004), reducing the frequency of switching to a minority of the population, controlling the pace of antigenic diversity generation (MacGregor et al. 2011), and potentially contributing

to the survival of parasites expressing alternative VSG coats below the threshold for immune detection (Gjini et al. 2010).

The virulence amongst *T. congolense* 'savannah' sub-type strains is mostly dependent on whether they are domestic or sylvatic, with highly virulent strains being more common in the sylvatic transmission cycle (van den Bossche et al. 2011). The increase in virulence may be the result of the selective pressure caused by the evolution of trypanotolerance in wildlife and may explain why the worst epidemics of AAT are seen following the introduction of foreign breeds into wildlife enzootic areas, as seen in Mozambique and Ethiopia at different times (Sigauque et al. 2000; Ayisheshim et al. 2015). The prevalence maintenance of low virulence strains in sylvatic cycles of areas with mixed virulence patterns has been explained by the cross-protection between low and high virulence strains shown in mice (Masumu et al. 2006). Within the domestic cycle, virulence seems to be highly strain dependent, irrespective of geographic area and genetic subgroup (Masumu et al. 2006).

The cross-protection given by low virulence strains may be susceptible to antigenic variation. The VSG repertoire in *T. brucei* is comprised of hundreds of pseudogenes, which constantly form new variants. Pseudogenes contribute to the pool of potentially expressed VSGs since they may form temporary mosaic VSGs by assembling with highly homologous intact genes (Marcello & Barry 2007a). This mechanism potentially confers the ability of the parasite to re-infect previously immune hosts. Furthermore, the generation of novel VSGs by recombination of the existing repertoire may offer a dynamic way of virulence and transmission increase through preservation of the parasite at subclinical levels and a greater host range.

1.11 Aims of the thesis

Despite being a major contributor to parasite's fitness, efforts to study antigenic diversity have been limited. Population genomics studies of African trypanosomes [in *T. brucei* (Sistrom et al. 2014; Weir et al. 2016) and *T. congolense* (Tihon et al. 2017)] have ignored VSG diversity perhaps because existing read mapping methods cannot be safely applied to labile VSG loci. Yet, VSGs and related genes are intimately linked to disease biology, particularly host range and virulence (Pays 2006). The size, complexity and dynamics of the VSG repertoire have challenged the development of tools for antigenic diversity and expression from big data. The purpose of this thesis was to develop a bioinformatics tool that allows the quantification and characterisation of the VSG repertoire from high-throughput data. Using next-generation sequencing (NGS) to produce genomes of natural African isolates, and transcriptomes from experimental tsetse fly infections, this thesis aimed to further characterise the VSG repertoires of *T. congolense* and *T. vivax* on a population scale, and to introduce the Variant Antigen Profile (VAP) as a metric of VSG diversity and expression. The VAP can define genetic signatures of parasite isolates, allowing a fast and reliable analysis of VSG diversity and expression, and bypassing gene annotation or individual VSG manipulation. As the VAP brings the unprecedented ability to discriminate among variant antigens, it will ultimately allow the association of variant antigens with specific infection phenotypes. This will trigger exceptional epidemiological mapping of *T. congolense* and *T. vivax*, leading to a better understanding of disease outcomes. Furthermore, I present putative VSG expression sites for *T. congolense* and compare evolutionary pressures in the molecular evolution of VSGs. The specific objectives of this thesis are:

1. Develop the VAP, a tool to analyse *T. congolense* VSG diversity on a large scale from genomic and transcriptomic data.
2. Apply the VAP to transcriptomic and proteomic data from experimental fly infections to characterise the mVSG repertoire of *T. congolense* 'savannah' Tc1/148.
3. Identify and describe the *T. congolense* VSG ES using long-read genome sequencing.
4. Extend the VAP to analyse antigen diversity in *T. vivax*.
5. Understand the role of recombination in generating VSG diversity in *T. brucei*, *T. congolense* and *T. vivax* at a population scale.

Chapter 2. The Variant Antigen Profile in *Trypanosoma congolense*: quantifying the frequency of conserved VSG motifs

2.1 Introduction

Studying antigenic diversity in African trypanosomes is challenging due to the 'dynamic nature' of antigenic variation of VSG sequences. Antigenic variation consists on the serial replacement of the cell surface protein, and prevents an effective antibody-based response, resulting in infection chronicity and increased likelihood of transmission, and hampering vaccine development. Antigenic variation is a key determinant both of disease progression and virulence; hence good knowledge of its effectors is required to understand how disease is caused and how to prevent it. In fact, describing the diversity of major surface antigens has been critical to understanding disease dynamics in other antigenically-variable organisms. For instance, antigen mapping of hemagglutinin has improved the prediction of antigenic drift in influenza A (McHardy & Adams 2009), a vital step for the efficacy of the seasonal influenza vaccine. Similarly, understanding of the genetic diversity of HIV envelope glycoproteins preceded the formulation of a multivalent subunit vaccine (McCutchan et al. 1996). In *P. falciparum* malaria, studies of *var* gene diversity have exposed the association between particular *var* genes and disease severity (Chen et al. 2011; C. W. Wang et al. 2012).

In *T. brucei*, the VSG repertoire has been known for a long time to be large, dynamic, and under fast evolution as a result of recombination events, such as telomeric exchange and gene conversion, and sexual reproduction (Myler 1993). However, VSG diversity in *T. congolense* and *T. vivax* is known to be generated differently (Jackson et al. 2012). For example, while the *T. brucei* VSG repertoire is thought to be effectively unlimited, in *T. vivax* it might be limited as strains expressing similar VSG repertoires can be found further apart and infection can provide cross-protection between close isolates (Nantulya et al. 1986). Yet, *T. brucei* remains the current model for antigenic variation and diversity. Most of what is known about the biology of *T. congolense* has derived from *T. brucei* studies, but

major differences in the VSG machinery, evolution and structure strongly advocate the need for dedicated study of this species. Much remains unknown about how parasite epidemiology, virulence patterns, clinically relevant genotypes, or population structure relate to VSGs. This suggests that in-depth study of antigenic diversity in *T. congolense* and *T. vivax* is necessary to achieve a full understanding of the molecular aspects of antigenic variation and reveal their biological significance.

The ancestor of African trypanosomes had two structurally different VSG families, a-type and b-type (Jackson et al. 2012). *T. brucei* has retained and expanded both types, but *T. congolense* only has b-type VSGs (Jackson et al. 2012). All VSGs have a NTD and a CTD, each characterised by different patterns of cysteine residues (Carrington et al. 1991). In *T. brucei*, three different types of NTD and six different types of CTD have been described (Carrington et al. 1991; Marcello & Barry 2007a). In b-type VSGs, the CTD has a conserved tertiary structure, indicating that VSG dimerization occurs via the NTD (Chattopadhyay et al. 2005). *T. brucei* b-VSGs derive from a bottleneck evolution from a single ancestral lineage arising after speciation (Jackson et al. 2012). Furthermore, the same NTD can be linked to different CTDs, showing that the NTD and the CTD are decoupled and indicating that the CTD is the downstream anchor point for recombination (Marcello & Barry 2007a). As all VSGs share a homologous CTD, recombination constraints are low, resulting in a VSG phylogeny of genome-specific long branches with narrower distributions (Jackson et al. 2012). This is a genome-specific feature as is observed in both a- and b-VSGs. Therefore, although the two lineages are structurally distinct, they share similar antigenic variability mechanisms (Jackson et al. 2012).

The scenario is different in *T. congolense*. At publication, the VSG repertoire of the *T. congolense* reference genome, IL3000, contained 875 genes. However, this is underestimated as 182 new VSG sequences have been added since. The repertoire contains only b-type VSGs and is organized into two structurally different families (Fam16 and Fam13), each combining multiple ancestral b-type VSG lineages (Jackson et al. 2013). The N terminus of a VSG is highly variable, but the tertiary protein structure remains conserved due to highly conserved amino acids at defined positions, mainly cysteine, glycine and tryptophan residues. Unlike *T. brucei*, the CTDs are distinct and can be used to differentiate VSGs into at least 15 distinct clades or 'phylotypes' (**Figure 9**). In *T. congolense*, the CTD is defined as the region up to 100 amino acids long between a conserved tryptophan residue and the GPI-

anchor. They are variable in length and composition, although, unlike the NTDs and the GPI-anchor, generally rich in polar residues. There is a clear compositional and structural difference between the CTD of Fam13 and Fam16, the latter being more homogenous. The degree of variation within the families is also distinct: the CTD of Fam13 phylotypes are more conserved than those of Fam16, which reflects higher genetic distances within the former.

The VSG phylogeny of *T. congolense* suggests wider distances between phylotypes, wider distribution of sequence variation, and the conservation of different ancestor lineages information within basal nodes. The lineage-specific CTD indicate that *T. congolense* VSGs lack a common recombination anchor point and the robust cladistic structure indicates little recombination among members of different phylotypes so the phylogenetic signal can be retained. This was confirmed by VSG quartet analysis, which supported scarce recombination between clades (Jackson et al. 2012). The lower VSG recombination rate proposed for *T. congolense* in comparison to *T. brucei* is also supported by the much lower occurrence of pseudogenes (30 % vs. 70 %), likely to result from gene conversion events.

Besides antigenic variation, VSGs and VSG-like sequences are important disease markers (Ross et al. 1987; Tetley et al. 1987; Allred et al. 1990; Carrington et al. 1991; Schwede et al. 2015) and have been implicated in disease biology, particularly host range and virulence (Pays 2006), as well as resistance to complement-mediated lysis (Ferrante & Allison 1983; Devine et al. 1986) and stimulation of cytokine dysregulation among innate immune cells, which ultimately leads to immune suppression and disease symptoms (Vincendeau & Bouteille 2006). Examples of VSG-like sequences that have evolved invariant roles are the serum resistance associated (SRA) gene in *T. brucei rhodesiense*, or the *T. brucei gambiense*-specific glycoprotein (TgsGP), both of which provide resistance to the human trypanolytic factor and therefore allow human infection (Van Xong et al. 1998; Uzureau et al. 2013; Capewell et al. 2013), as well as the transferrin receptors (TFR), which allow iron uptake into the cell in both *T. brucei* and *T. congolense* (Salmon et al. 1997), and ESAG2, a *T. brucei*-specific gene family derived from b-type VSGs, which have an unspecified function in the flagellar pocket of bloodstream-form *T. brucei* (Gadelha et al. 2015). The conservation of distinct lineages within the *T. congolense* VSG repertoire is consistent with negative selection, which might be driven by functional differentiation of particular clades.

The intimate relationship between VSGs, virulence, and pathology further strengthens the hypothesis that these phylotypes may hold important biological functions. Yet, population genomic studies to date have not been able to address VSGs, perhaps because current read-mapping methods cannot accurately handle such great variability in copy number, sequence composition and chromosomal localisation. Characterisation of hundreds of VSGs, including many novel mosaics and pseudogenes, is a daunting process and has so far only been undertaken for reference genomes (Marcello & Barry 2007a; Jackson et al. 2010; Jackson et al. 2012; Hall et al. 2013). There is an urgent need for a systematic classification capable of dealing with the complexity of VSG repertoires, but simple and sensitive enough to be applied to large sets of samples of variable genome quality. Manual analysis of the antigen repertoire is arduous, time consuming and requires specific expertise in VSG recognition and sequence manipulation. The high percentage of pseudogenes in the repertoire takes this challenge even further. However, the predictability of the *T. congolense* VSG repertoire facilitates the development of a tool to quickly screen a genome and estimate its relative VSG repertoire composition. Here, I introduce the VAP as an automated method to quantify pathogen antigenic diversity and define antigenic signatures of parasite strains, allowing a fast and reliable analysis of *T. congolense* VSG variation from genome sequence data which can potentially impact disease phenotypes or strain virulence, and trigger unprecedented epidemiological mapping of *T. congolense* in Africa.

This chapter aims to:

1. Investigate the universality of the VSG phylotypes identified in the reference strain using phylogenetic analysis;
2. Present VAPs for a historical collection of 41 field strains from across Africa using a manual method based on sequence similarity search to quantify population variation in the VSG repertoire;
3. Introduce the VAP as a method to dissect antigenic diversity in *T. congolense*;
4. Quantify VSG variation at the population level and link it to geography and population structure.

2.2 Methods

2.2.1 Sample identification and storage

Samples were collected between 1966 and 1982 by International Livestock Research Institute (ILRI, Kenya) staff from livestock, tsetse flies and wild animals as part of various experiments. They were passaged into a naïve host to increase parasitaemia. At peak parasitaemia, naïve hosts were bled and multiple 50µl blood stabilates in 10-30 % glycerol prepared and stored at -80C. Samples were transferred to liquid nitrogen at a later date. Samples for this project were chosen from the ILRI Biobank based on area of collection and number of passages. All samples used had had a maximum of 4 passages, as passaging through multiple hosts may affect the VSG repertoire. Samples were imported from ILRI during 2014 and 2015 and kept in liquid nitrogen at the Biosciences Building, University of Liverpool (**Table 4**).

2.2.2 Cell lysis and DNA extraction

Blood stabilates were subject to magnetic antibody cell sorting to deplete host cells and enrich for parasite DNA using anti-CD15 and anti-CD45 antibodies, as most leukocytes have one or both of these antigens. DNA was extracted using a magnetic bead protocol. The magnetic bead: sample ratio was adjusted to 2:1 to increase recovery of small DNA.

150 µl blood samples were transferred to 2.0 ml Lo-Bind tubes (Eppendorf, UK) and 1.5 ml ACK lysing buffer (0.15 M NH₄Cl, 10 mM KHCO₃, 0.1 mM EDTA) was added. Samples were incubated for 3 min at room temperature, and centrifuged for 10 min at 650 g. Supernatant was discarded and the pellet washed in 500 µl MACS buffer (2mM EDTA, 5xBSA in PBS pH7.2). Samples were centrifuged for 10 min at 650 g. Supernatant was discarded and 80 µl MACS buffer, 10 µl antiCD15, and 10 µl antiCD45 were added, gently mixed until pellet dissolved, and incubated at 4 °C for 15 min with gentle shaking after 7.5 min. Cells were washed in 1 ml MACS buffer and centrifuged for 10 min at 650 g. Supernatant was discarded and pellet resuspended in 500 µl MACS buffer. MACS columns were washed with 500 µl MACS buffer and cell suspension was added to individual columns. The first eluate was collected and centrifuged for 10 min at 650 g. Supernatant was discarded and

the pellet resuspended in 100 μ l lysis buffer (aqueous solution of 1 M Tris-HCl pH8.0, 0.1 mM NaCl, 10 μ M EDTA, 5 % SDS, 0.14 μ M Proteinase K).

Samples were incubated at room temperature for 1 hour and DNA was extracted with magnetic Sera-Mag Speedbeads (GE Healthcare Life Sciences, UK). Samples were transferred to 1.5 ml LoBind tubes (Eppendorf, UK), two volumes of previously vortexed, room temperature Sera-Mag Speedbeads were added, and incubated for 15 min at room temperature. Tubes were placed on magnetic stand for 2 min and the supernatant was removed. 500 μ l of freshly made 70 % ethanol was added, incubated for 2 min and removed. The latter step was repeated to maximise DNA precipitation. All visible ethanol was carefully removed and the pellet air-dried for 5 min or placed on a 37 °C heat block for 3 min. The pellet was rehydrated in 20 μ l nuclease-free water and placed on the magnetic stand for 1 min. The supernatant was transferred to a sterile 0.5ml LoBind tube (Eppendorf, UK) and kept at -20 °C after Qubit® fluorometric dsDNA quantitation (dsDNA HS Assay Kit) (Life Technologies, UK).

DNA outputs obtained ranged from 2.0 ng to 57.2 ng.

Table 4 *T. congolense* strains used in this study. Table shows sample ID, year and location of collection, host, and species used for passaging.

Sample ID	Year	Location	Host Species	Passage Species
IL3900	1980	Bobo Upper Delta, Burkina Faso	Dog	Mice
IL3926	1980	Bobo Upper Delta, Burkina Faso	Dog	Mice
IL3897	1982	Bobo Upper Delta, Burkina Faso	Cattle	Mice
IL2995	1983	Bobo Upper Delta, Burkina Faso	Bovine	Rat
IL3578	1983	Bobo Upper Delta, Burkina Faso	Bovine	Rat
IL3932	1992	Delmonte, Kenya	Horse	Goat
IL274	1976	Kabete, Kenya	Dog	Mice
IL374	1976	Kabete, Kenya	Dog	Mice
IL399	1976	Kabete, Kenya	Dog	Mice
IL409	1976	Kabete, Kenya	Dog	Mice
IL410	1976	Kabete, Kenya	Dog	Mice
IL438	1976	Kabete, Kenya	Dog	Mice
IL439	1976	Kabete, Kenya	Dog	Mice
ILC55	1976	Kabete, Kenya	Dog	Mice
ILC66	1976	Kabete, Kenya	Dog	Mice
IL3688	1982	Kenya	Lion	Mice
IL3686	1982	Kenya, Kenya	Lion	Mice
IL3035	1985	Muhaka, Kenya	Bovine	Mice
IL396	1976	Nairobi, Kenya	Dog	Mice
IL3779	1991	Nguruman, Kenya	Tsetse fly	Mice
IL3296	1972	Robanda, Tanzania	Bovine	Mice
IL1180	1961	Serengeti, Tanzania	Lion	Mouse
ILC22	1970	Serengeti, Tanzania	Bovine	Rat
IL2068	1971	Serengeti, Tanzania	unknown	Mouse
IL3949	1972	Taita, Kenya	Bovine	Bovine
IL311	1979	The Gambia	Bovine	Rat
IL3674	1979	The Gambia	Bovine	Rat
IL3675	1979	The Gambia	Bovine	Rat
IL2992	1966	Transmara, Kenya	Bovine	Mice
IL3019	1966	Transmara, Kenya	Bovine	Mice
IL3021	1966	Transmara, Kenya	Bovine	Mice
IL3022	1966	Transmara, Kenya	Bovine	Mice
IL3180	1966	Transmara, Kenya	Bovine	Mice
IL3349	1966	Transmara, Kenya	Bovine	Mice
IL3775	1966	Transmara, Kenya	Bovine	Rat
IL2326	1962	Uganda	Bovine	Mice
IL588	1962	Uganda	Bovine	Mice
IL3978	1992	unknown	Bovine	Mice
IL3304	1967	Zaria, Nigeria	Bovine	Mice
IL3954	1967	Zaria, Nigeria	Bovine	Mice
IL2281	1979	Zaria, Nigeria	Bovine	Mice

2.2.3 Magnetic antibody cell sorting control

Success of the magnetic antibody cell sorting was observed by PCR to amplify the mouse 18S rDNA region and the ITS1 rDNA region of *T. congolense*. PCR was

performed in a reaction volume of 25 µl with the following conditions: Bioline 2X Taq mix (Bioline, London), 2 µM of each forward and reverse primers, and 1mM dNTPs. The reaction conditions were as follows: 1 cycle of 94 °C for 3 min followed by 30 cycles of 94 °C denaturation for 1min, 58 °C hybridisation for 1 min, and 72 °C elongation for 1 min, and a final extension step of 5 min at 72 °C.

The primers used were as follows:

- Mouse 18S rDNA primers: 5'-GTAACCCGTTGAACCCCAT-3' and 5'-CCATCCAATCGGTAGTAGCG-3'
- *T. congolense* ITS1 rDNA 5'-GCG TTC AAA GAT TGG GCA AT-3' and 5'-CGC CCG AAA GTT CAC C-3' (Desquesnes et al. 2001).

PCR products were resolved by gel electrophoresis (1h at 100V) in 1.5 % agarose gels stained with 1:10,000 dilution in TBE of SYBR safe DNA Gel stain (Thermo Fischer Scientific, UK) and visualised under UV light.

2.2.4 Genomic amplification

To maximise the chance of success through genomic library preparation, genomic amplification using the illustra GenomiPhi V2 DNA Amplification Kit was performed to all samples according to the manufacturer's protocol (GE Healthcare Life Sciences, UK). Amplification products were quantified with Qubit® fluorometric dsDNA quantitation (dsDNA Broad Range Assay Kit) (Life Technologies, UK).

Genome amplification products from 11 samples were used for library preparation because the library preparation procedure failed to yield sufficient material from the direct DNA extraction (samples IL274, IL374, IL396, IL409, IL439, IL3022, IL3296, IL3775, IL3900, IL3954 and ILC22). The remaining products were stored at -20C.

2.2.5 Next-Generation Sequencing (NGS)

To proceed with genomic sequencing, genomic libraries were prepared using the Accel-NGS 2S DNA Library Kit (Swift Biosciences, Inc., USA). This library kit was chosen because it is designed to produce effective yields from DNA inputs as low as 10pg, includes sequential repair steps and produces high complexity libraries, which

are required for successful deep sequencing. Furthermore, it uses a magnetic bead protocol for library purification similar to that used in the DNA extraction step. Protocol was followed as per manufacturer's instructions and is comprised of 6 steps: DNA was mechanical sheared using the Covaris ADA process (Covaris, UK) in 500 bp fragments, followed by two repair steps and two ligation steps. Libraries were amplified by PCR using 10-18 cycles.

Following preparation, DNA library concentrations were measured with Qubit® fluorometric dsDNA quantitation (dsDNA Broad Range Assay Kit) (Life Technologies, UK). Genomic libraries with detectable DNA levels were bioanalysed using a High Sensitivity DNA Analysis Kit (Agilent Technologies, USA). Samples with desirable concentration ($>0.15 \text{ ng.}\mu\text{l}^{-1}$) and fragment lengths (300-1000 bp) were kept. Samples with different fragment lengths (N = 15) were subject to Pippin preparation (Pippin Prep, Sage Science, USA) for targeted size selection, as recommended for Illumina sequencing.

Genomic libraries were sequenced on a MiSeq platform (Illumina Inc, USA) as 150 or 250 bp paired ends, at the Wellcome Trust Sanger Institute (Cambridgeshire, UK). The whole genome shotgun project has been deposited in SRA under the accession number ERP023223.

2.2.6 Analysis of NGS data

De novo assembly

Draft genomes were produced for each strain so that VSG-like sequences could be recovered. *De-novo* assembly was manually optimised for the dataset (see Results section 2.3.2).

Recovered reads were used to produce *de novo* assemblies of each strain using velvet version 1.2.07 (Zerbino & Birney 2008) with the following settings: kmer of 65, insert length adjusted for 400 base pairs with standard deviation of 50, minimum pair count of 20, and coverage cut-off between 0 and 5, depending on sample quality.

VSG-like sequence recovery

To recover all VSG-like sequences in the genomes, a sequence similarity search was performed. Since the genetic distance between strains was unknown and VSG are dynamic genes, it was difficult to predict the degree of similarity to expect between VSGs. Therefore, identity thresholds were kept low to ensure any novel or atypical VSGs were recovered.

Assembled contigs were examined for VSG-like sequences by sequence similarity search with tBLASTx using a database of *T. congolense* IL3000 VSG as query. Although IL3000 lacks a-type VSGs (Jackson et al. 2012), a separate sequence similarity search was performed using *T. brucei* Fam1 and a-type VSGs as query to confirm the absence of a-VSGs in this dataset.

A threshold of p-value > 0.001, contig length >150 amino acids, and % identity >=75 was applied to select significant results. Sequences with 40 to 75 % similarity to the reference were manually inspected and their inclusion in the analysis empirically decided. This search would allow for the recovery of VSG sequences that were absent from the reference, yet no novel phylotypes were identified in the analysis. Recovered sequences were assigned a phylotype based on their top hit from the reference and phylotype relative frequencies used to estimate manual VAPs. VAPs are defined as the proportions of each VSG phylotype for a given strain.

Multiple sequence alignment and phylogenetic estimation

VSG-like nucleotide sequences were manually retrieved from the assembled contigs files, translated with BioEdit 7.2.5 (<http://www.mbio.ncsu.edu/BioEdit/bioedit.html>), and aligned to a subset of the IL3000 VSG database with ClustalW (EBI, UK) (Larkin et al. 2007).

Fam16 and Fam13 multiple sequence alignments and phylogenies were produced for each strain to investigate whether the cladistic structure of the IL3000 VSG phylogeny was conserved throughout the species. For each strain, a phylogeny was estimated from a protein sequence alignment of recovered VSG-like sequences and IL3000 VSG sequences with the neighbour-joining (NJ) method and the WAG+ Γ substitution model (Whelan and Goldman 2001) using MEGA7 (Kumar et al. 2016). Each phylogeny was manually investigated for the presence of novel phylotypes,

which would be reflected in particular clades or branches of strain VSGs, with no orthologue in IL3000.

All full-length VSG sequences from IL3000, IL3675 (The Gambia), and IL3900 (Burkina Faso, Forest sub-type) were translated to amino acid and aligned with ClustalW (EBI, UK) (Larkin et al. 2007) to produce a VSG phylogeny representative of the *T. congolense* species. The alignment contained 1037 sequences in total, 778 from IL3000, 214 from IL3675, 31 from IL3900, 12 from *T. brucei* ESAG2 and 2 from *T. vivax* b-type VSG as the outgroup. The representative VSG phylogeny was estimated from protein sequence alignments with the maximum likelihood (ML) method and the WAG+ Γ substitution model (Whelan and Goldman 2001) using RAxML v.2 (Stamatakis 2014) and PHYML (Lefort et al. 2017) following amino acid model selection in PHYML (Lefort et al. 2017). Robustness was assessed with 100 bootstrap replicates. Bayesian inference (BI) trees were estimated with gamma rates function in MrBayes v3.1.2. (Huelsenbeck & Ronquist 2001; Ronquist & Huelsenbeck 2003) and two Markov chain Monte Carlo chains run in parallel over 5,000,000 generations, with a burnin of 2,500 and a fixed WAG+ Γ model to ensure the program would reach stationary convergence. Posterior probabilities of each node were used to assess accuracy of BI trees. Neighbour-Joining (NJ) trees were estimated in Phylip (Felsenstein 1989) using the executables protdist and neighbour for a multiple dataset of 100 bootstrap replicates and a random seed of 99.

To evaluate the robustness of each phylotype and whether the allocation of particular VSGs to each clade was significant, maximum likelihood ratios of each clade were calculated and compared with RAxML (Stamatakis 2014) in triplicate. For each strain and clade, a VSG was randomly chosen and forced to cluster in the adjacent clade and negative log-likelihood of unconstrained trees was compared to that of constrained trees.

To investigate the population structure and strain relationships of the dataset, A SNP phylogeny was estimated from SNP alignment with ML method under the JTT+ Γ substitution model with RAxML v.2 (Stamatakis 2014). Robustness was assessed with 500 bootstrap replicates. For details on SNP calling process, see section 2.2.7.

Phylotyping

The representative VSG protein sequence alignment was used to identify unique strings of 9-59 amino acids that could serve as diagnostic markers for each phylotype. Twenty-eight motifs were identified through consecutive VAP estimation simulations using a positive control of reference VSGs. This includes 593 sequences from all 15 clades, which are captured by 28 amino acid motifs (**Figure 10**). These motifs were included as query to each assembled contigs file of the field strains in HMMER under default parameters (Eddy 1998) and their relative frequency used to create the automated VAP.

Figure 10 shows the consensus sequence of the motifs used and their distribution through the VSG primary structure. The number of motifs required to unequivocally capture VSGs of a certain phylotype is dependent on the genetic distance between phylotypes and sequence variation level within the phylotype. For example, phylotype 10 contains three groups of genetically distant sequences, which have to be targeted specifically. On the other hand, genetically homogenous phylotypes, such as phylotype 1, 5, 6, 7 and 12, only required a single motif. There is also the case of very large phylotypes, such as phylotype 15, which although homogenous and with a defined CTD sequence, required multiple motifs. In this situation, the size of the phylotype introduces a higher number of polymorphisms. The motifs, although very similar to each other, must be searched for individually so that all sequence variations can be identified and accounted for.

To compare the stability of the composition of the VAPs to the background random variation, the total pool of VSGs recovered in the study was used to create 41 randomised simulated VAPs based on 250 VSGs each.

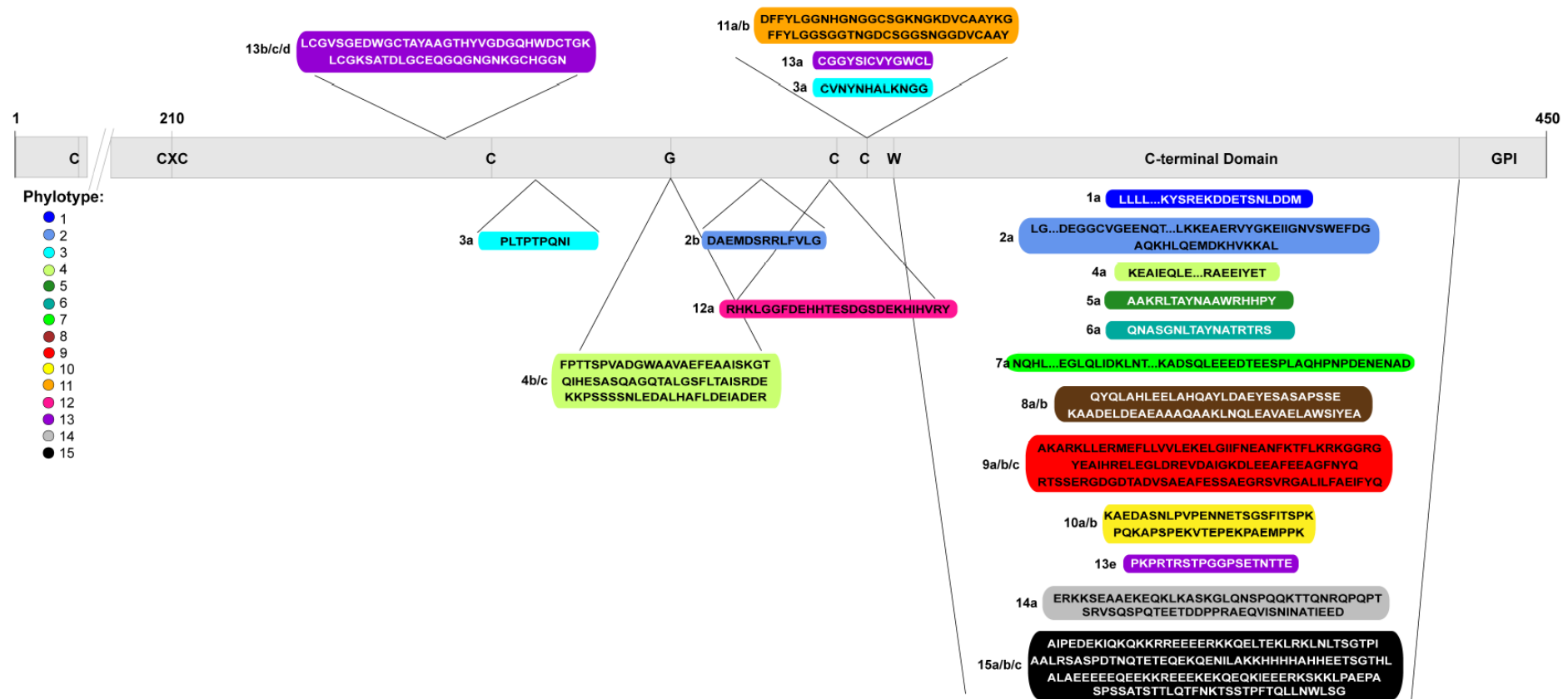


Figure 10 Diagnostic motif consensus sequences of 15 *T. congolense* VSG phylotypes and their positioning in the primary amino acid structure (1 to 450 amino acids). Phylotypes are colour-coded according to key. Amino acid sequences of the protein motifs are shown for each phylotype at the corresponding location on the VSG primary structure.

2.2.7 Strain variation

To produce an alignment comparing single nucleotide polymorphisms (SNPs) across the whole genome of the sample cohort, MiSeq reads were retrieved and mapped against the *T. congolense* IL3000 genome using BWA mem (Li 2013), converted to bam format, sorted and indexed with Samtools (Li et al. 2009). Sorted bam files were cleaned, duplicates marked and then indexed with Picard (<http://broadinstitute.github.io/picard/>). Single Nucleotide Polymorphisms (SNPs) were called and filtered with Genome Analysis Toolkit (GATK) suite according to the recommended protocol for multi-sample variant calling (Van der Auwera et al. 2013). Reads were realigned and loci called with GATK (Van der Auwera et al. 2013). SNPs were called for individual samples, but genotypes were called simultaneously for all samples. Finally, SNPs were extracted as a multi-sample file and filtered using default parameters using GATK (Van der Auwera et al. 2013). The multi-sample vcf file obtained from GATK was converted to fasta format using vcftools v0.1.14 (Danecek et al. 2011) and a maximum likelihood phylogeny was estimated with RAxML, using the JTT+ Γ model of nucleotide substitution, following nucleotide model selection on MEGA7 (Kumar et al. 2016).

2.2.8 Statistical analysis

The statistical comparisons between the BLAST and the VAP performances in recovering VSGs were done using the Pearson's correlation test in R (RStudio Team 2016). Outliers were identified using a threshold of $2x\sigma$ with the function 'removeOutlier' in R (RStudio Team 2016). Outliers were manually inspected before removal as described in section 2.3.6. F-tests were performed to analyse variance between observed and simulated data.

2.3 Results

2.3.1 Antibody-based cell sorting efficacy

The CD45 antigen, formerly called leukocyte common antigen, is a protein tyrosine phosphatase, receptor type C, exclusively expressed by leukocytes. To maximise parasite DNA recovery from the blood stabilates, CD45+ cell depletion was performed (**Figure 11** and **Figure 12**).

Figure 11 shows fluorescence-activated cell sorting of CD45+ cells before and after CD45+ cell depletion using magnetic anti-CD45 antibodies on an MACS column. The post-treatment graphs (right) show a clear reduction of the immunofluorescent CD45+ population, suggesting a successful depletion. **Figure 12** shows the result of the CD45+ depletion on DNA recovery by PCR of host and parasite markers. Host DNA was amplified with mouse 18S RNA primers and parasite DNA with ITS1 primers. The absence of a band in lane 2 shows that mouse DNA was successfully depleted. The presence of a band in lane 7 suggests that some parasite DNA is lost during the procedure, remaining in the mouse cell fraction. However, this loss has less impact in the sequencing results than a high ratio of host: parasite DNA would have.

With the evidence that host cell depletion can be successfully achieved using magnetic antibody cell separation, CD15+ antibody was also added to increase depletion success. CD15+ is a cluster of differentiation antigen widely expressed in phagocytic leukocytes, particularly in neutrophils. The addition of this antibody results in a higher depletion chance, therefore contributing to better parasite fraction enrichment.

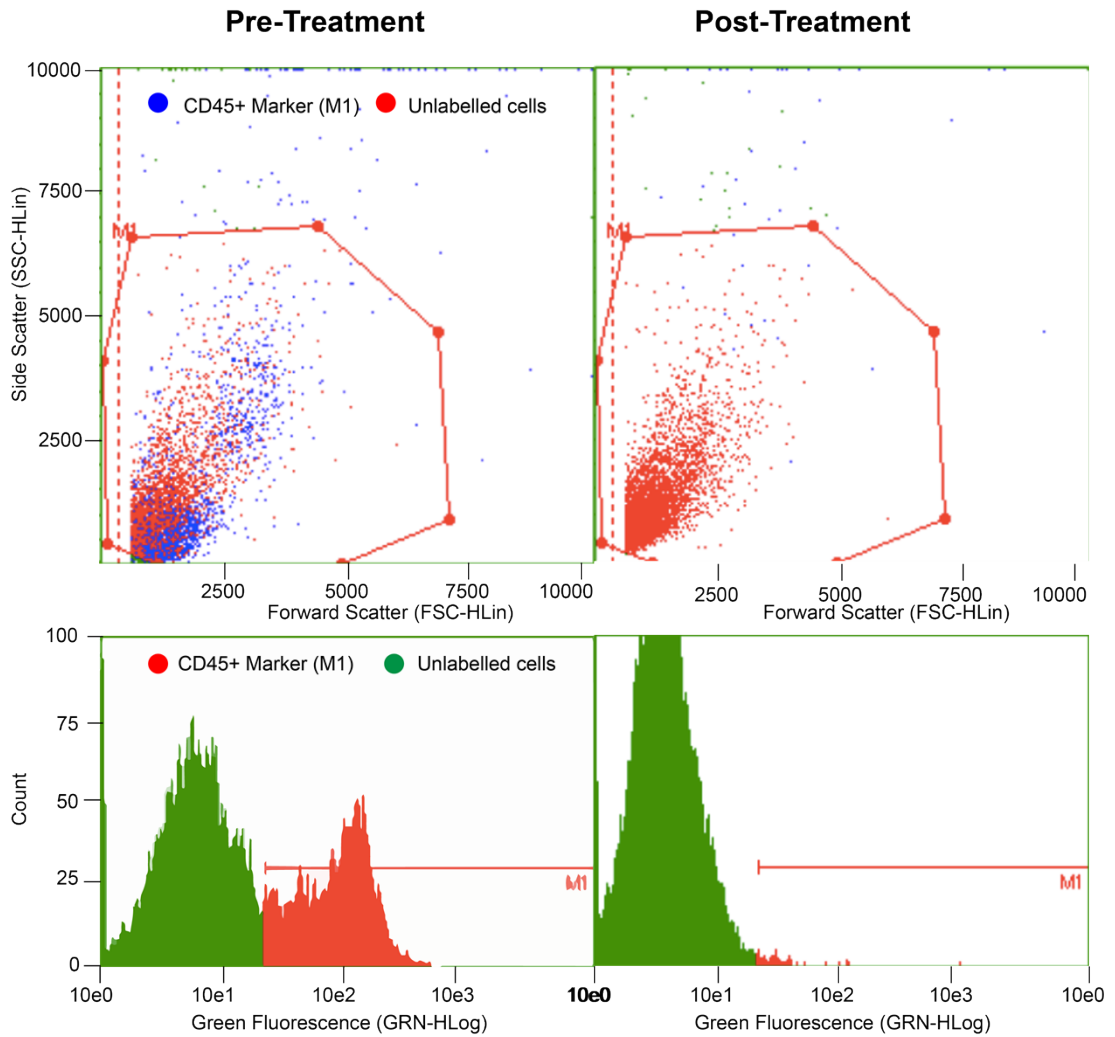


Figure 11 Fluorescence-activated cell sorting of CD45+ cells. Whole mouse blood infected with *Trypanosoma congolense* was lysed with ammonium chloride and leukocytes were stained for CD45. Pre-treatment refers to original blood sample, post-treatment refers to eluate after CD45+ cell depletion using magnetic beads column. CD45+ population in marker M1 (bottom) is backgated onto forward scatter (FC)/ side scatter (SS) in blue (top). (Noyes, H., unpublished).



Figure 12 Gel electrophoresis UV picture of PCR products from CD45+ cell depletion. Lane 1: Mouse 18S rRNA primer, before depletion. Lane 2: Mouse 18S rRNA primer, parasite fraction after depletion. Lane 3: Mouse 18S rRNA primer, Mouse cell fraction after depletion. Lane 4: Mouse 18S rRNA primer, negative control. Lane 5: *T. congolense* ITS1 primer, before depletion. Lane 6: *T. congolense* ITS1 primer, parasite fraction after depletion. Lane 7: *T. congolense* ITS1 primer, Mouse cell fraction after depletion. Lane 8: *T. congolense* ITS1 primer, negative control. Marker: Bioline 100 bp ladder. (Noyes, H., unpublished).

2.3.2 *De novo* assembly

Sequencing of the clinical isolates resulted in an average of 3.02×10^6 paired reads per sample, ranging between 6.19×10^5 and 1.33×10^7 .

To find the best *de novo* assembly software for the dataset, three programs (i.e. Velvet, SOAPdenovo2, Abyss) were tested against four strains under default conditions and kmer of 65 (**Figure 13**). Optimal assemblies produce long contigs, which is reflected on low total contig number, high N50 and high maximum contig length. Whilst occasionally the Abyss assembly had a higher n50 (IL438), Velvet 1.2.10 (Zerbino & Birney 2008) provided the best balance of all parameters. Abyss (Simpson et al. 2009) assemblies resulted in higher contig number, whereas SOAPdenovo2 (Luo et al. 2012) performed poorly overall.

After choosing Velvet, different kmers were tested in three strains to identify the optimal assembling conditions for the dataset. Kmers between 49 and 75 were tested in intervals of 5 until the maximum N50 could be established. A kmer of 65 was found to yield the best assembly results and therefore was used for subsequent

assemblies (**Figure 14**). *De novo* assemblies of the clinical isolates had an average n50 of 803 with a range of 161 to 3257.

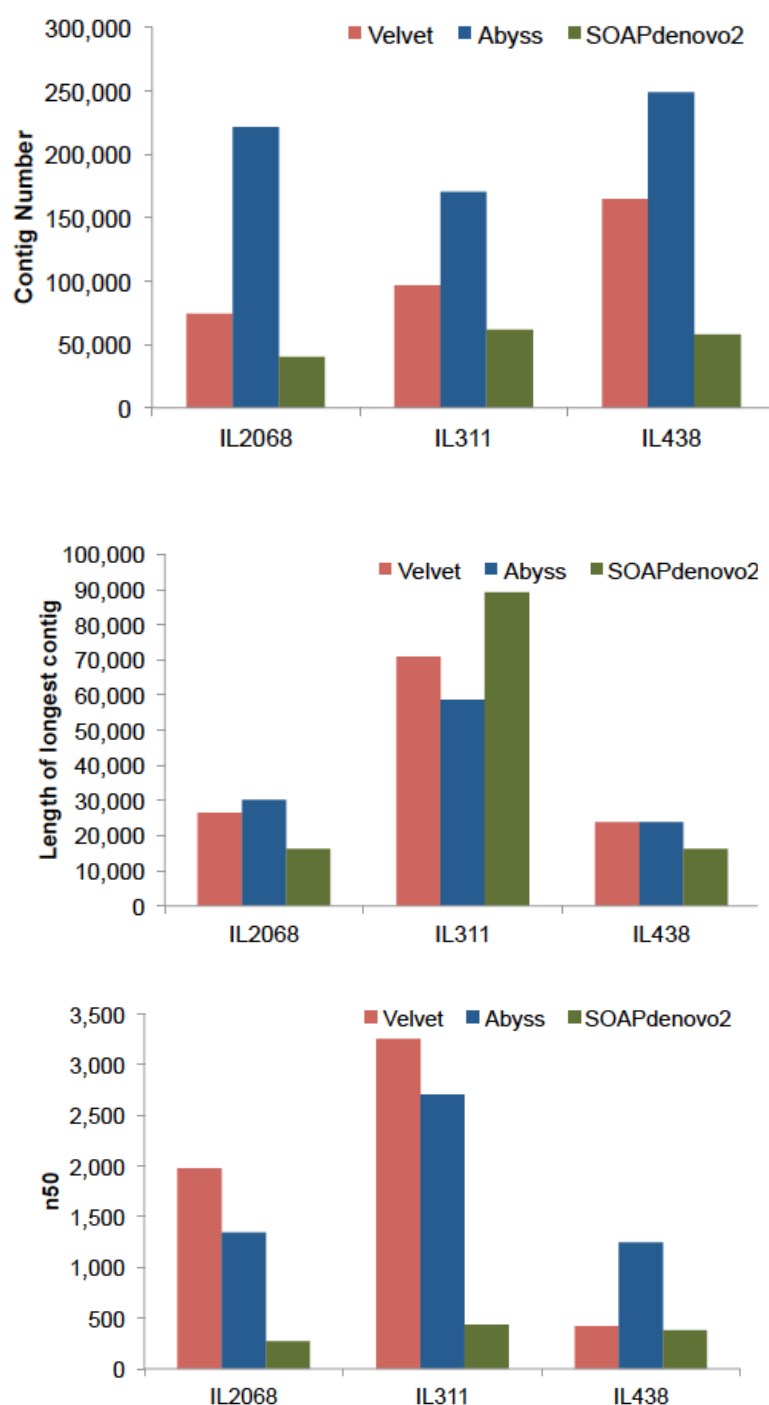


Figure 13 Comparison of performance *de novo* assemblers. Velvet, Abyss and SOAPdenovo2 were compared, using N50, contig number and maximum length of contigs as measured in three different *T. congolense* strains.

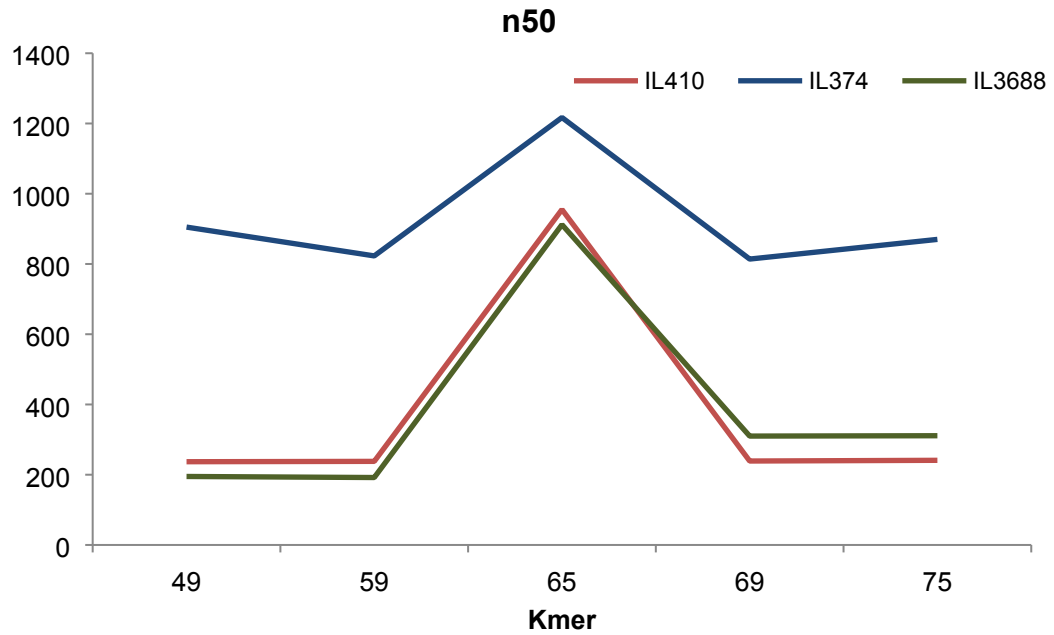


Figure 14 Assembly quality comparison with different kmer values, using Velvet 1.2.10 (Zerbino 2010). Three strains (IL410, IL374, and IL3688) have been used in the comparison. With the exception of the kmer, all assembly parameters were kept constant.

2.3.3 Sampling test

Obtaining the complete repertoire of VSGs from a field sample is unlikely due to the frequently low parasitaemia of natural infections. To understand whether profiling could be accurately done with partial repertoires, a sampling test was performed on the IL3000 VSG repertoire. Random selections of increasingly smaller VSG numbers were subject to manual profiling. Although the variant antigen profile may be affected by the quality of the sample, the relative proportions of each phylotype are maintained with as little as 20 % of the repertoire. Simulations of various profiles of the reference IL3000 with decreasing sampling numbers showed good correlations to the original (**Table 5**). This indicates that samples containing at least 20 % of VSG repertoire can still be accurately profiled.

Table 5 Correlation analysis results (partial vs. full repertoire VAP). r = Pearson's moment correlation.

Repertoire %	r	Adjusted R^2
75	0.998	0.995
50	0.978	0.9521
30	0.975	0.9462
20	0.948	0.8917
10	0.893	0.7825

2.3.4 Phylotype universality and phylogeny revision

For each strain, a VSG phylogeny was estimated to investigate if all reference phylotypes were represented and if novel phylotypes existed. All sequences from field strains placed at the terminal nodes of the phylogeny with the reference sequences and all phylogenies maintained the cladistic structure observed in IL3000. As a representative of the species VSG structure, a phylogeny combining VSGs of the reference (IL3000, Kenya), a West African strain (IL3674, The Gambia) and 'forest' sub-type strain (IL3900, Burkina Faso) was estimated to confirm that no strain-specific clades, absent from IL3000, were observed. These isolates were chosen because they incorporate the maximum amount of variation due to their high geographical and genetic distance. The topology of this consensus VSG phylogeny is similar to the reference phylogeny, showing that the phylotypes identified in IL3000 are present in every strain and that the internal relationship between phylotypes is conserved (**Figure 15**). No novel phylotypes were identified.

Log-likelihood ratio tests were performed to confirm that all VSG sequences could be robustly placed within established phylotypes. For each VSG, the log-likelihood of an unconstrained tree was compared to another in which the VSG was constrained within the adjacent clade. Log-likelihood ratio tests were conducted for randomly sampled VSGs of IL3675 and IL3900, in triplicate for each clade. The negative log-likelihood of unconstrained trees was significantly higher than constrained trees in all cases ($p < 0.01$), except for phylotype 2 of IL3900. This was significantly different from the adjacent phylotype 1, but not from the adjacent

phylotype 3. These tests confirm that the 15 VSG phylotypes seen in *T. congolense* IL3000 are robust.

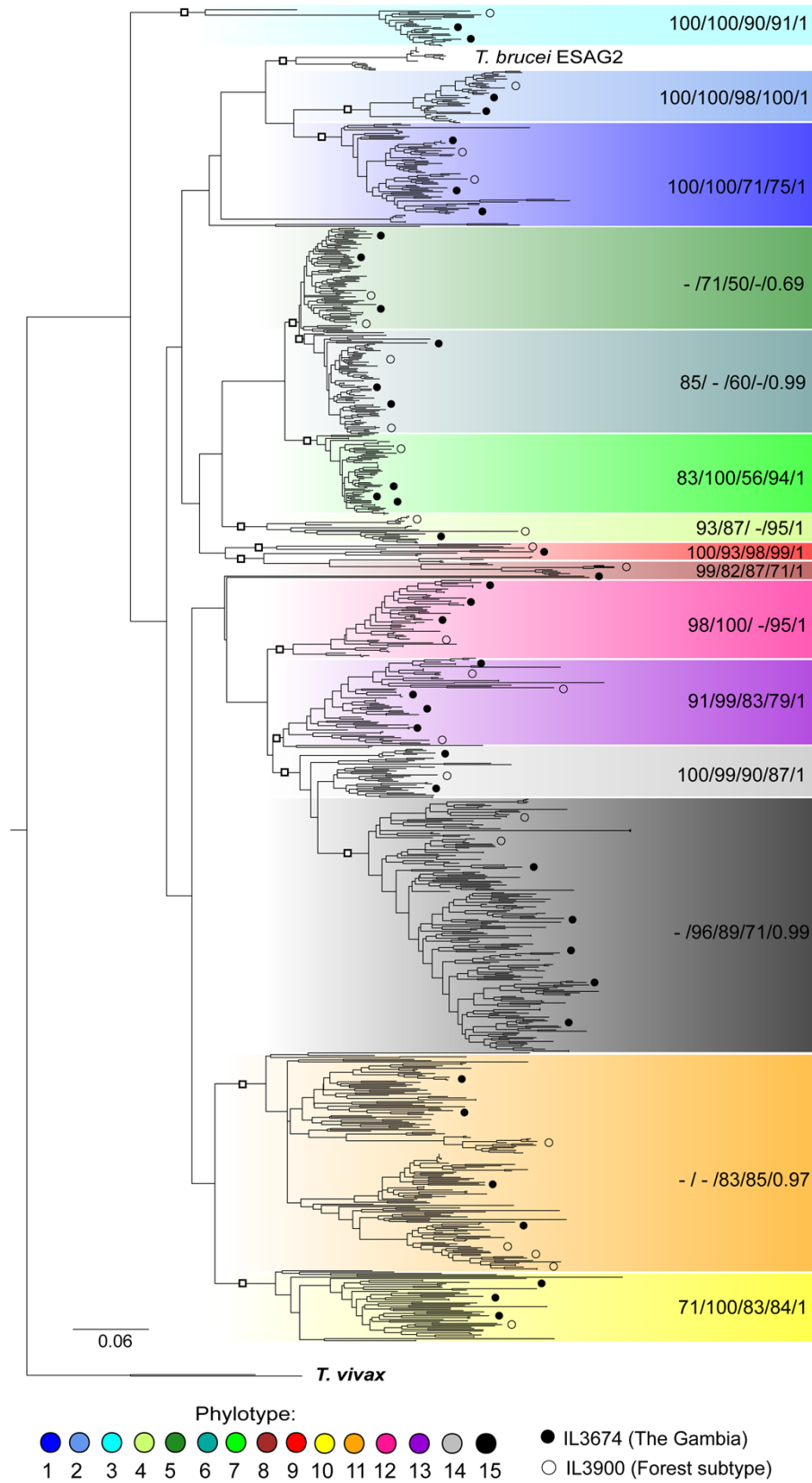


Figure 15 Maximum likelihood phylogeny of *T. congolense* full-length VSG using IL3000 (Kenya), IL3674 (The Gambia), and IL3900 (forest sub-type). The phylogeny was estimated from protein sequences with RAxML (Stamatakis 2014) using a maximum likelihood method with a WAG+Γ model and 100 bootstrap replicates. The fifteen phylotypes identified in IL3000 are colour-coded according to key. Position of example sequences from IL3674 and IL3900 are indicated according to key. Internal nodes are labelled with bootstrap percentages for maximum likelihood (ML) from the complete tree (RAxML), and ML (PhyML) (Guindon et al. 2010), ML (MEGA7) (Kumar et al. 2016), neighbour joining (NJ) (Felsenstein 1989), and posterior probabilities (BI) (Huelsenbeck & Ronquist 2001) estimated from a pruned tree containing 147 sequences. Tree is rooted with two *T. vivax* VSG sequences (Fam23).

2.3.5 Positive Control

Unique structural motifs were designed to distinguish between phylotypes and obtain a VAP from assembled contigs. An automated system for motif search was developed and the designed motifs were tested on a positive control set composed of 593 IL3000 VSG sequences from all phylotypes. The positive control shows a very high correlation with the full IL3000 repertoire ($R^2 = 0.97$, Pearson's product moment correlation, $t(13) = 21.98$, $p < 0.001$). The motif search correlates well with the manually-curated profile obtained by both manual phylogenetic localisation and sequence similarity search ($R^2 = 0.98$, Pearson's product moment correlation, $t(13) = 379.77$, $p < 0.001$), showing that motif search can recover phylotype proportions that are known.

2.3.6 Sequence similarity vs. structural motif searches

Sequence similarity and structural motif-based profiles were produced for each strain. Correlation analyses were performed to compare the two outputs. The motif search recovered more VSGs than the sequence similarity search (mean $\pm \sigma = 721 \pm 277$ vs. 669 ± 292 , paired t-test, $p = 0.005$) and individual phylotype correlations were often poor. A closer analysis of the data revealed outliers in particular phylotypes, these were manually inspected and removed (from this stage of the analysis only) and correlations re-estimated (**Table 6**).

Table 6 Correlation values for individual phylotypes before and after improvement (Pearson's Moment Correlation) and description of the outliers removed.

Phylotype	Pearson's <i>r</i>		Strains removed
	Before Improvement	After Improvement	
1	0.71	0.81	IL274
2	0.51	0.81	IL2281
3	0.57	0.77	IL2281
4	0.31	0.64	IL2326, IL3779, IL1180, IL2281, IL2068
5	0.79	0.79	
6	0.63	0.73	IL311, IL2281
7	0.84	0.88	IL3900, IL396
8	0.38	0.75	IL3932, IL2281, IL1180, IL1769, IL3779
9	0.02	9.59	IL3949, IL3932, IL2068, IL3900, IL1769
10	0.39	0.51	ILC55, IL3932, IL3775, IL2281
11	0.72	0.80	IL2281
12	0.57	0.70	IL3779, IL1769, IL3900, IL3897
13	0.48	0.79	IL3900, IL396, IL3779, IL399
14	0.62	0.75	IL3779, IL588, IL3932
15	0.74	0.81	IL2281, IL3900
Total	0.81	0.82	

The initial comparison between methods showed a very high correlation to the manually-curated IL3000 repertoire ($R^2 = 0.88$, Pearson's product moment correlation, $t(13) = 9.7321$, $p < 0.001$) (**Figure 16, left**), but a weak correlation to the full dataset. Several outliers were identified and closer analysis revealed one strain with recurrently poor results (i.e. IL2281 with 8 out of 15 datapoints being outliers), whereas the remaining outliers were sporadic and from different strains. Reasons for the sporadic mismatch between the sequence similarity search and the motif search seem to be related to both sequencing data quality and filtering thresholds in the sequence similarity software. In instances where the sequence similarity software recovered more VSGs than the motif search, these were often due to an individual VSGs being divided between 2 contigs, resulting in the software counting every partial VSG as an individual entry, meaning that BLAST finds high identity in a contig containing the N terminus and then the same in the contig containing the CTD. In strains with poorer assembly, i.e. shorter contigs and fewer full sequence VSGs, this effect is accentuated.

In instances where the motif search recovered more VSGs than the sequence similarity search (e.g. IL2281), the cause was often the opposite. Unusually long contigs with multiple VSGs result in the sequencing similarity search software

allocating those contigs based on the best matching VSG (ignoring the other). In contrast, the motif search allocates them according to the presence of the structural model for every appropriate phylotype. This way, a single contig containing multiple VSGs and therefore multiple diagnostic motifs will be placed multiple times, whereas BLAST will consider it a single entity and allocate only one single similarity hit.

The removal of 40 outliers (6.5 %) resulted in an improvement of individual phylotype correlations (**Table 6**) and a good positive correlation in the full dataset (**Figure 16, right**) ($R^2=0.67$, Pearson's product moment correlation, $t(566) = 34.39$, $p < 0.001$), supporting the use of the motif search for profiling. Besides recovering the reference strain VSG repertoire, our motif-based method is more sensitive to VSG sequences than a BLAST-based annotation.

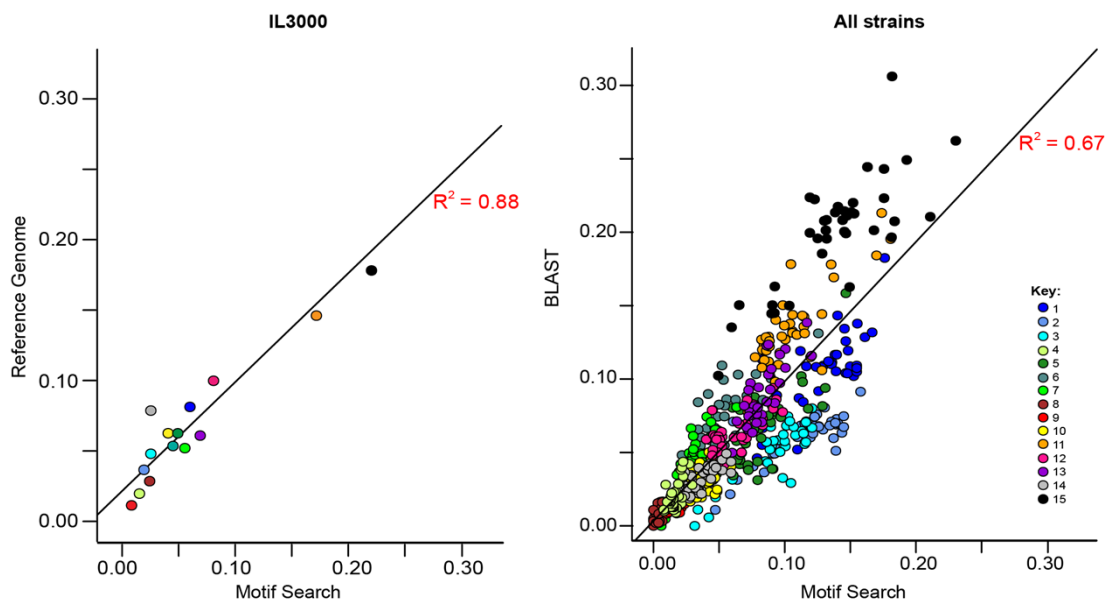


Figure 16 Performance of the protein motif-based VAP compared to the manual estimation of phylotype proportion in IL3000 and all the field isolates. Manual VAPs were estimated by counting the results from sequence similarity search [BLAST (Altschul et al. 1990)]. The adjusted R^2 statistics showing the correlation of both methods is indicated in red. Phylotypes are colour-coded according to key.

The Variant Antigen Profiling pipeline has been compiled in a Perl script hosted in the Galaxy server (<https://usegalaxy.org/>) and available to download for local installation from GitHub (<https://github.com/ssilva1/VAPPER>). The pipeline

processes raw genomic sequencing data of a single isolate and produces a table of phylotype frequencies, heatmaps of phylotype abundances and variations, and a PCA plot putting the sample in the context of the available *T. congolense* genomic isolates. This currently includes our dataset and the isolates published by Tihon et al. (2017).

2.3.7 Antigenic diversity in *T. congolense*

The global scale of VSG gene diversity is commonly thought to be very large, similar to that observed for *P. falciparum* var genes due to the effect of immune selection. To examine this issue, we compared the VAPs in our *T. congolense* sample set and observed that the composition of the VAP is stable across *T. congolense* isolates (**Figure 17, heatmap**). Phylotypes 1, 11 and 15 are consistently the most numerous, whereas phylotypes 4, 8, and 9 consistently scarce (**Figure 17, bar chart**). To assess whether the stability observed was statistically significant, the observed frequencies were compared to 41 simulated VAPs, each estimated from 250 VSGs, randomly selected from the pool of VSG retrieved for all samples combined. The simulated VAPs showed significantly more variation in relative phylotype proportions than strain genomes (F-test, $p < 0.001$) (**Figure 18**), showing that the stability of the genomic VAPs is significant.

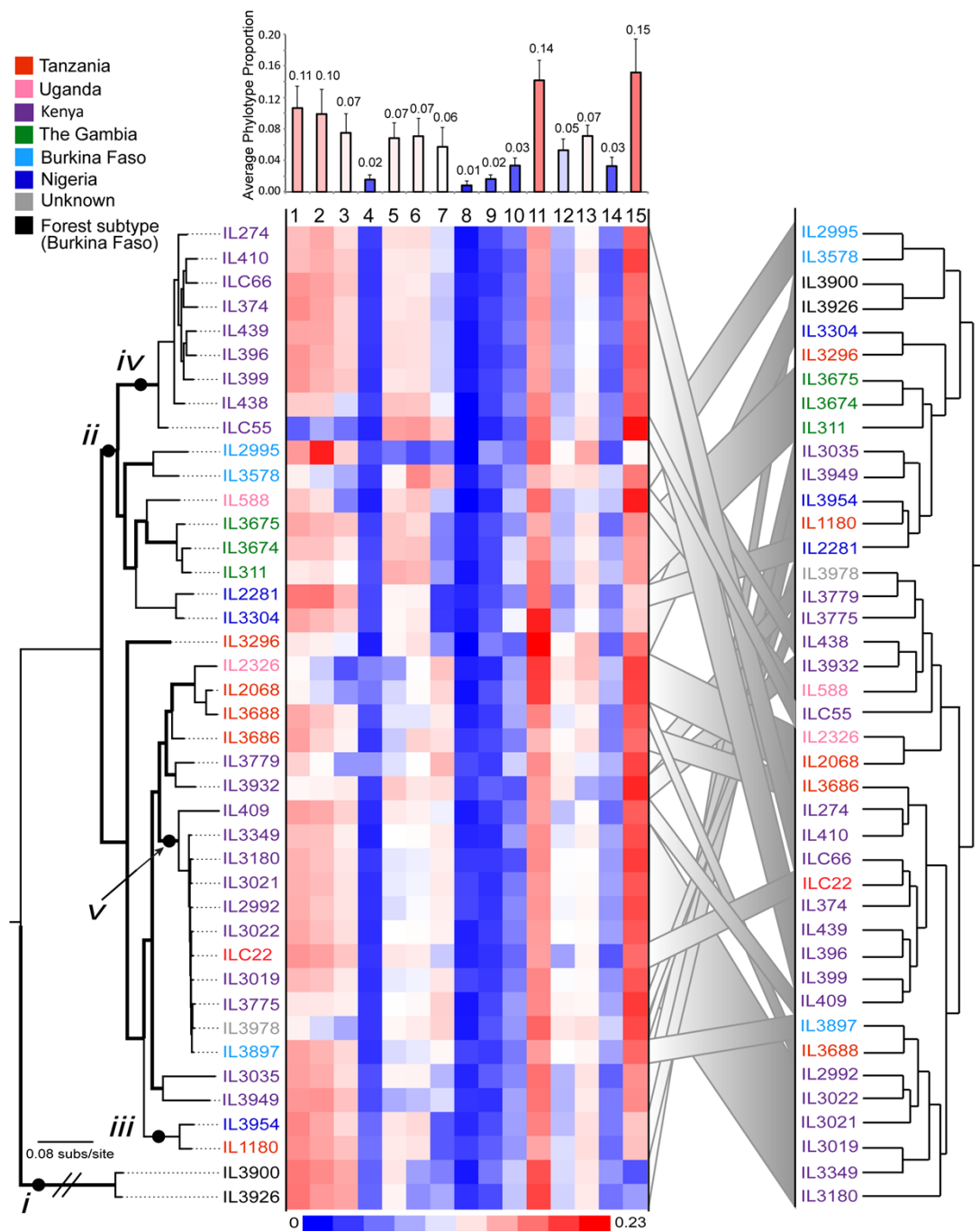


Figure 17 The relationships between the VSG repertoire, geography and population structure in *Trypanosoma congolense*. On the left, the strain relationships are described by a ML phylogeny estimated from whole-genome SNP data, using RAxML (Stamatakis 2014) with a GTR+Γ model and 100 bootstrap replicates, which are shown as thick branches. The heatmap describes the phylotype proportion across the strain genomes. VAPs are organised according to the SNP phylogeny. On the right, a dendrogram illustrates the VAP relationships within the population. Strains are colour coded by location of collection according to key. Labels *i* to *v* are referred in the text. Average phylotype proportions (mean+σ) are shown at the top.

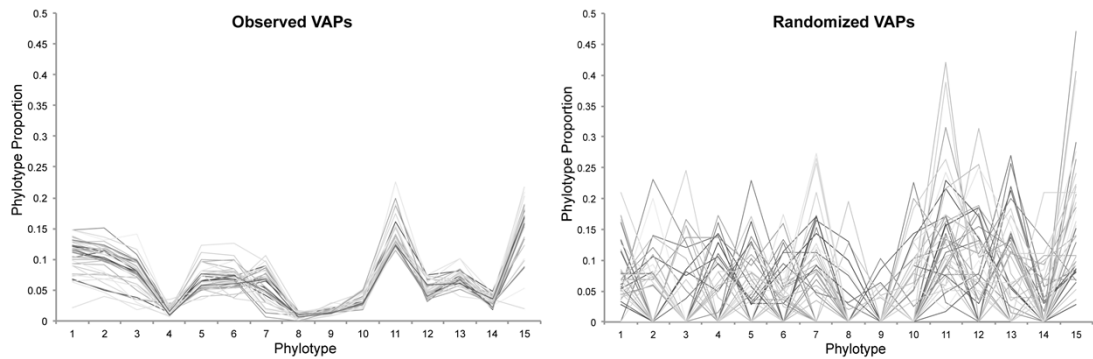


Figure 18 Variation in phylotype proportions estimated from the real (left) and the randomized (right) data (N = 41 for each condition). Simulated VAPs were estimated from random selections of 250 VSGs from a pool of all VSGs recovered in this study. The observed VAPs show less variation than the randomized data (F-test, p-value<0.001).

To compare how the VAP correlates to the whole genome variation, SNPs were called and converted into a sequence alignment to produce a phylogeny of the all strains included in the dataset (**Figure 17, left**). The tree shows a clear separation between ‘savannah’ and forest sub-types (denoted by ‘i’). Within the ‘savannah’ sub-type, there is a clear geographical signature only on the top part of the phylogeny (IL274 to IL3304, denoted by ‘ii’), where strains separate according to country of collection. The remaining isolates lack a geographical signature, particularly when looking at the short phylogenetic distance between IL3954 and IL1180 (denoted by ‘iii’), from Nigeria and Tanzania, respectively. This suggests that these isolates represent at least two, but possibly more subpopulations of *T. congolense* ‘savannah’, highlighting the complexity of the *T. congolense* population structure and the need for extensive sampling in order to accurately represent it.

There is a tenuous association between the VAP and the population structure. When comparing VAPs of genetically close isolates, conserved patterns can be seen for some, but not all groups. For instance, among seven Kenyan samples (IL274 to ILC55, denoted by ‘iv’), there is very little variation in SNPs. Yet, ILC55 has a very distinctive VAP, specifically due to the abundance of phylotypes 5-7 and the scarcity of phylotypes 1-3. ILC55 clusters with IL3932, IL588, IL2326 and IL2068, three of which are from a different population group and were isolated in different countries (i.e. Kenya, Uganda and Tanzania). Another example is ILC22 from Tanzania, which is genetically close to the Kenyan strains IL3349 to IL3775

(denoted by 'v'), but whose VSG repertoire is similar to the Kenyan strains depicted in 'iv', particularly due to lower abundance of phylotypes 7 and 12.

Whilst labelling and sample handling errors are a possibility when working with historical isolates, these differences may indicate a genuinely weak association between the VAP and both geography and population structure, which might suggest that genetic variation segregating in the sub-telomeres is decoupled from core chromosomal loci. Importantly, this could suggest that other factors are involved in inheritance of the VSG repertoire and consequently in how antigenic diversity is generated in *T. congolense*.

Although the relative proportions of VSG phylotypes appear to be a fixed feature of the *T. congolense* genome, they are not entirely invariant. After normalisation by the cohort mean, fluctuations in phylotype size can be detected and signature patterns start to emerge (**Figure 19**). For example, a subset of samples from Kenya, Uganda, Tanzania and Burkina Faso (IL3978 to IL3578, denoted by 'i') show a signature of under-represented phylotypes 1-3, whereas the Gambian samples show over-representation of phylotypes 5 and 6 (denoted by 'ii'). Additionally, the 'forest' sub-type isolates show a distinct over-representation of phylotype 15 (denoted by 'iii'). Furthermore, the degree of variation in phylotype abundance is positively correlated to gene number itself, as high-abundance phylotypes express the highest variation (phylotype 11 and 15) and low-abundance phylotypes are less variable (e.g. phylotype 8 and 9) ($R^2 = 0.74$). This variation may reflect gene gain and loss within phylotypes on a population scale, which may have functional implications.

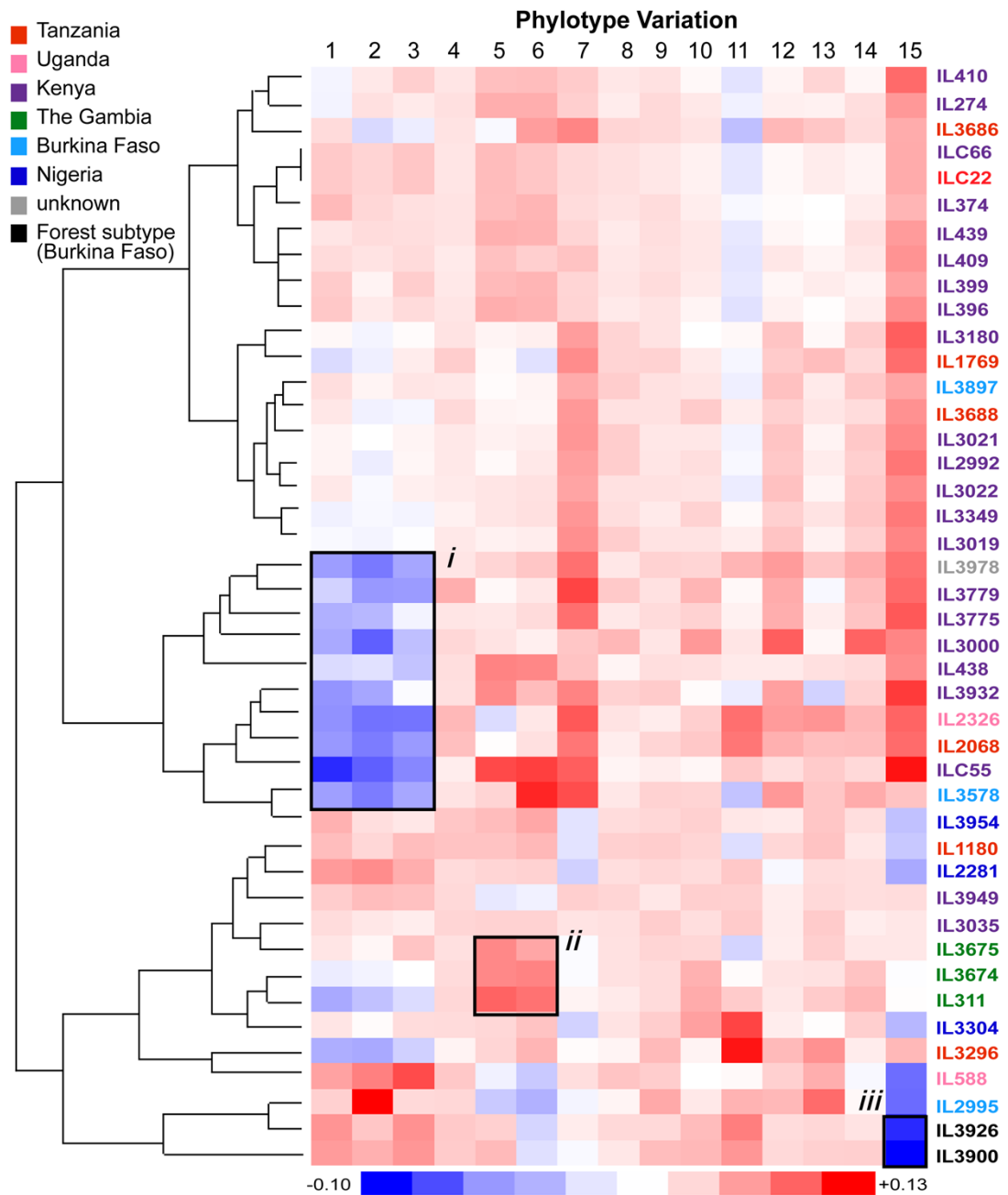


Figure 19 Phylotype variation across the population. The heatmap describes phylotype variation across the population expressed as the deviation from the mean. The dendrogram describes the relationships among the VSG repertoires of each strain. Strains are colour coded by location of collection according to key. Labels *i* to *v* are referred to in the text.

2.4 Discussion

2.4.1 Antigenic diversity in *T. congolense*

Previous research on *T. congolense* VSGs has been restricted to the reference strain IL3000; therefore, this study is the first to put VSG diversity in a population context. I have sampled 41 isolates from 6 countries that were collected at different times. 39 out of 41 isolates were passaged into rodents, which may have led to some diversity selection in the sample set because some *T. congolense* isolates may not expand in mice. Due to the low parasitaemia of natural trypanosome infections, this is inevitable when dealing with field isolates, at least until targeted sequencing is further developed. Here, I show that the 15 *T. congolense* VSG phylotypes proposed for IL3000 (Jackson et al. 2012) are present in a diverse collection of strains from both 'savannah' and forest sub-types. The absence of novel gene lineages among strain genomes suggests that the original phylotypes are exhaustive and capable of accommodating all VSG diversity observed to date. I have shown that their genomic proportions can be quantified using protein structural motifs. Hence, I believe that the VAP allows the high-throughput analysis of VSG repertoires from any *T. congolense* genome sequence.

The fact that the *T. congolense* VSG repertoire is self-contained and restricted to these 15 defined groups is a major disparity in VSG evolution between *T. brucei* and *T. congolense*. Sequences from similar phylotypes, but belonging to different strains, are more closely related to each other than to other sequences from the same strain but different phylotypes. This is consistent with the view of Jackson *et al.* (2012) that *T. congolense* VSG repertoire is a combination of multiple ancestral b-type VSG lineages with distinct CTDs, as opposed to *T. brucei*, where a and b-type VSG share a common CTD. Restricted recombination between phylotypes indicates opposite evolutionary pressures to maintain antigenic diversity, whilst conserving phylogenetic structure. This, together with differences in size and variance across the dataset, suggest these phylotypes may have distinct purposes, which empowers them as potentially diagnostic if related to pathology, virulence, or host use.

The VAP is important because understanding antigenic diversity is critical to appreciate disease distribution, virulence patterns, and immune evasion pathways,

as observed in many other organisms. For instance, in pathogenic *Neisseria* spp., the effector proteins of antigenic variation are linked to virulence (Helaine et al. 2007). Like VSGs, pilins are antigenically variable, heavily glycosylated, surface glycoproteins that exist in both truncated soluble and membrane-bound forms. The number of expressed pilins and their variation degree is strain-dependent and their main function is cell adhesion (Virji 2009). However, minor forms are also involved in other roles, such as environment adaptation, e.g. low-iron response and foreign DNA uptake (Virji 2009). The latter is crucial for increasing the transformation frequency of bacteria and thus sustaining genetic diversity (Fussenegger et al. 1997). Additionally, pilin complexes act directly on pilus fibres for movement retraction and the mechanical force generated may be involved in cell signalling and immune modulation by triggering the cortical plaque structure formation and the shedding of regulatory factor CD46, respectively (Winther-Larsen et al. 2005; Maier et al. 2002).

In the bacterial pathogen *Anaplasma marginale*, antigenic variation occurs due to variation of the Major Surface Protein genes (*msp*). There are five families of *msp* genes (*msp1-5*); they modulate infection, development of persistent infections, and transmission [reviewed by de la Fuente et al. (2005)]. Whilst *msp1* is a single-copy gene involved in cell adhesion (De la Fuente et al. 2003), Chávez et al. (2012) have shown that *msp2* expression pattern is dependent on the host cell type. Specifically, the number of *A. marginale* MSP2 variants was found to be different in various cell lines, experimental and natural infected tick cells and mammalian cells, and unique variants were found in each of the cell types. These data show that the composition of MSP2 variant populations changes according to host cell characteristics, and suggest MSP2 has a role in infection and survival ability in divergent hosts (Chávez et al. 2012). Recently, *Anaplasma phagocytophilum* was genotyped in natural endemic cycles by PCR amplification of 16S rRNA, groEL, *msp2*, and *msp4* genes (Frölich 2017). A relationship with continent of isolation was investigated and found for the antigenically variable *msp2* gene and groEL, but not for *msp4* genes, which instead separate by host. The geographical signal of *msp2* was also reported by Morissette et al. (2009) when analysing North-American strains of *A. phagocytophilum*.

In *Plasmodium falciparum*, var genes are the major surface antigens of the blood stage of the parasite. They play a role in establishing chronic infection and in promoting transmission to the mosquito, and they are thought to be functionally

differentiated (Chen et al. 2011). More importantly, *var* genes encode the *Plasmodium falciparum* erythrocyte membrane protein-1 (PfEMP1), an important player in virulence through the mediation of adhesion to host endothelial receptors and evasion of splenic clearance (Lavstsen et al. 2003). Analysis of the genetic diversity of *var* genes in clinical isolates of variable disease severity has uncovered a relationship between *var* gene subsets with different roles in host-parasite interactions (C. W. Wang et al. 2012). Expressed *var* gene repertoires of severe malaria isolates were compared to asymptomatic controls and isolates from severe malaria patients were found to preferentially transcribe group A *var* genes, a result that appears to be reproducible in Brazil, Kenya and Mali (Kirchgatter & Del Portillo 2002; Bull et al. 2005; Kyriacou et al. 2006). Wang et al. (2012) also found multiple unique *var* sequences in the severe malaria isolates only, which could potentially be associated with disease severity.

Var genes have also been linked to geography. Whilst Wang et al. (2012) found little overlap between *var* repertoires from Tanzania, Albrecht et al. (2010) found that the *Plasmodium* isolates circulating in Western Amazon had a limited *var* gene repertoire. In the largest epidemiological study of *var* genes to date, small scale spatial diversity was found, represented by repertoire conservation within each country, but little overlap between the repertoires of African isolates and even less between those from South America, Africa and Asia (Chen et al. 2011). The degree of diversity uncovered in local African populations was also found to be much greater than the previously reported diversity in Brazil or Papua New Guinea, which is consistent with wide genome diversity and microsatellite analysis (Anderson et al. 2000; Mu et al. 2005).

All these examples show that variant antigens are important players in disease progression and phenotype, disease distribution, transmissibility and virulence, but that they can also have functions other than antigenic variation. On that basis, such roles can also be predicted for VSGs.

In kinetoplastid research, the potential usefulness of studying VSG patterns to differentiate between strains and understand disease distribution have long been recognised (Meirvenne et al. 1977). In 2007, Hutchinson et al. reported for the first time that VSG repertoires evolved to become strain-specific. In a comparison between the *T. brucei gambiense* strain TbgDal isolated in Ivory Coast, the Kenyan *T. brucei brucei* strain Tb927, and strains collected from eight field isolates of the *T.*

b. rhodesiense-endemic district Tororo, in Uganda, it was found that the Ugandan strains had VSG homologues (greater than 40 % identity) of similar genetic distances in both East and West African *T. brucei* genomes, suggesting they all share a common ancestor. However, sequence divergence between homologues was found to be non-random with respect to protein structure. Also, sequence change was concentrated in antigenic regions, i.e. amenable to immunoglobulin recognition (those both solvent accessible and at the distal end of the VSG), suggesting an independent pressure for the creation of antigenically novel VSGs within each strain. In *T. brucei*, pathogenicity is partly strain-specific (Morrison et al. 2010), being strongly influenced by host: parasite interactions and requiring distinct host mechanisms. Specifically, Morrison et al. (2010) showed that infection with different strains results in different patterns of innate immunity, suggesting that innate immune response modulation is a major element in trypanosomiasis and greatly dependent on the parasite genetic background. Together these findings emphasise the importance of an integrated host-parasite approach that incorporates parasite diversity into future studies of pathogenesis.

Besides diagnostic potential, the results of this chapter provide some evidence of functional differentiation among the *T. congolense* VSG repertoire. Functional differentiation among *T. brucei* VSGs is well documented. For example, the TFRs (present in *T. congolense* also) have evolved from VSGs expressed at the parasite surface to become essential iron-binding proteins expressed at the flagellar pocket. TFRs have diverged since speciation, having expanded considerably in *T. congolense* and becoming VSG expression site associated proteins in *T. brucei* (Jackson et al. 2013).

The origins of human infectivity in some *T. brucei* strains provides examples of VSG-like genes evolving to non-variant functions by independent mechanisms (De Greef & Hamers 1994; Berberof et al. 2001; Van Xong et al. 1998; Uzureau et al. 2013; Capewell et al. 2013). In *T. brucei rhodesiense*, human infectivity is conferred by the expression of the SRA gene from the active VSG expression site. The SRA is a VSG-like gene positioned either in long tandem arrays with VSG genes or at a specific expression site. It contains 6 of the 8-conserved cysteine residues characteristic of *T. brucei* a-VSG, but has lost the surface-exposed antigenic loops (Blum et al. 1993; De Greef & Hamers 1994; Campillo & Carrington 2003; Vanhamme et al. 2004). When *T. brucei rhodesiense* comes into contact with human serum, which is rich in trypanolytic factors (TLF-1 and 2) associated with

expression of the apolipoprotein L1 (APOL-1) protein, the expression site containing SRA becomes active (Van Xong et al. 1998). The SRA protein is targeted to the endolysosomal compartment where it directly binds and inactivates APOL-1 (Vanhamme et al. 2003).

In *T. brucei gambiense*, human infectivity has evolved independently, but also from a truncated VSG-like sequence (Capewell et al. 2013; Uzureau et al. 2013). TgsGP is a VSG-like sequence lacking the VSG CTD and located in the sub-telomeric region of chromosome 2 (Capewell et al. 2013; Uzureau et al. 2013). Unlike SRA, TgsGP does not directly interact with APOL-1. In contrast, TgsGP induces membrane stiffening and haptoglobin-hemoglobin (Hp-Hb) receptor inactivation. This prevents TLF-1 to enter the parasite, resulting in acceleration of APOL1 degradation (Uzureau et al. 2013). These examples in *T. brucei* provide basis for the prediction that in *T. congolense* particular VSGs or VSG phylotypes may also have evolved novel roles within the complexities of host-parasite interactions.

Following these examples in *T. brucei*, the segregation of *T. congolense* VSGs into 15 conserved clades seems consistent with functional divergence within the repertoire of this species. This is reinforced by the lack of recombination between phylotypes (Jackson et al. 2012) and the evidence for them being under purifying selection comparable to the genomic background. The average value for ω (d_n/d_s) among orthologous VSGs in different strains is 0.27 (N = 1034). With the exception of phylotypes 8-10, which show a more neutral substitution rate (0.73; N = 123), this is not significantly different from the average ω for single-copy orthologues across the genome (0.19; N = 694; $p > 0.05$). This indicates that VSG primary structure does not evolve especially fast, but rather is strongly conserved, and further suggests that recombination provides the sole means of VSG diversification in *T. congolense*.

As the VAP appears to be a fixed feature of *T. congolense* 'savannah' (and is not substantially altered in 'forest' sub-type either), it is possible that these phylotypes have different functions. Under the hypothesis that phylotypes are functionally redundant, existing to increase VSG structural diversity, I would expect a trace of neutral or positive evolution. In both cases, phylotype sizes in the genome would vary randomly across the sample set due to gene gains and losses that were selectively neutral. Such process would result in low-abundance phylotypes (e.g. 8) being sporadically absent. Instead, the results show that the proportions of each

phylotype vary little across the sample cohort and that variation is lower in low-abundance phylotypes. This signature, consistent with negative selection, strongly suggests that VSG phylotypes are maintained in specific proportions because they fulfil distinct roles.

2.4.2 Strain relationships with space and time

When investigating the patterns of phylotype variation, East and West African samples do not separate clearly. When these findings are put into the context of whole genome variation, it becomes clear that even at the genomic level East and West African samples of the ‘savannah’ sub-type are not two mutually exclusive populations of *T. congolense*. This contrasts with the epidemiology of human-infective *T. brucei*, which fall into geographically-defined populations: the western *T. brucei gambiense* and the eastern *T. brucei rhodesiense* (Gibson 2001). Instead, *T. congolense* forms two clades comprising strains from various locations. In only one clade (**denoted by “ii” in Figure 17**) samples cluster geographically. Clearly, there is a danger that this may reflect inadequate and non-systematic sampling. However, the addition of data from another population genetic study of 52 *T. congolense* strains by Tihon *et al.* (2017), confirms that variation in VSG profile is not coupled to population history, and only partially explained by geography (**Figure 20**).

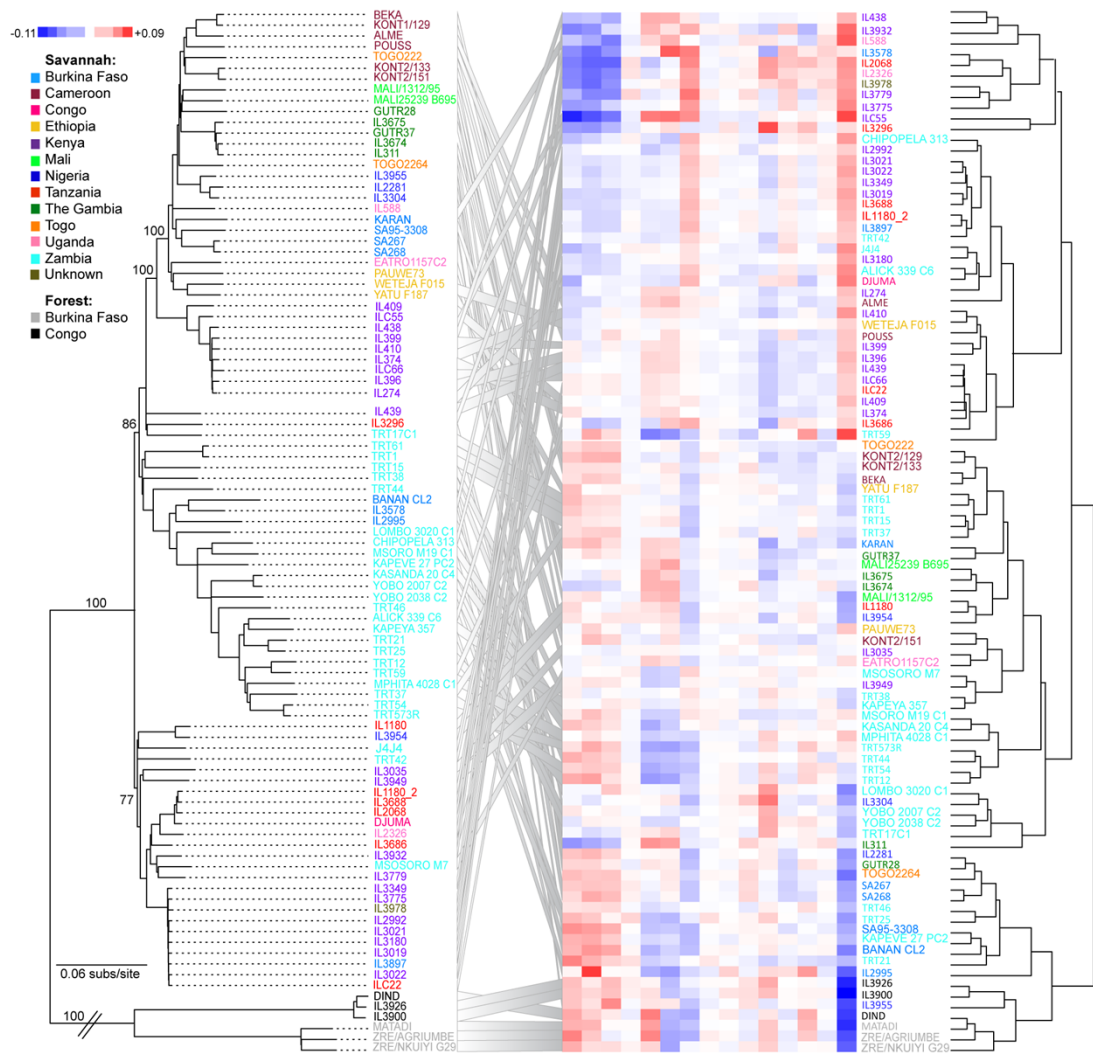


Figure 20 The relationships between the VSG repertoire, geography and population structure in *Trypanosoma congolense* including the isolates published by Tihon *et al.* (2017). On the left, the strain relationships are illustrated by the whole genome single nucleotide polymorphisms maximum likelihood phylogeny. The phylogeny was estimated with RAxML (Stamatakis 2014) with a GTR+ Γ model and 100 bootstrap replicates, which are shown as branch labels. The heatmap illustrates the phylotype variation across the cohort expressed as deviation from the mean. VAPs are organised according to the cluster analysis shown in the dendrogram on the right. Strains are colour coded by location of collection according to key.

The uncoupling of VSG repertoire and population history suggest at least three different interpretations: first, that *T. congolense* ‘savannah’ has genetically distinct populations that precede the species spreading across Africa, resulting in multiple populations circulating in the same region; second, that it derives from cattle movement across Africa, triggering parasite migration and diversification; third, that

it results from mating, as the hemizygous subtelomeres telomeres would be inherited in a non-Mendelian fashion during meiosis. The latter is supported by evidence for sexual reproduction in *T. congolense* (Morrison, Tweedie, et al. 2009; Tihon et al. 2017). In Gambian *T. congolense* 'savannah', a high level of diversity with a large number of distinct genotypes and a population division in four sub-populations was reported (Morrison, Tweedie, et al. 2009). Tihon et al. (2017) provided robust confirmation of genetic recombination amongst *T. congolense* populations by showing evidence of diploid hybrid parasites circulating in natural populations in Zambia.

Indeed, the literature confirms that *T. congolense* has a complex population structure. For example, microsatellite loci from Cameroonian *T. congolense* 'forest' sub-type vary greatly in size and are not strictly structured (Simo et al. 2013; Simo, P. S. Fogue, Melachio, et al. 2014). Population sub-structuring in *T. congolense* is consistent with the observations in other parasites, specifically the small scale geographic structuring in African *P. falciparum* (Chen et al. 2011) and *T. brucei* (Maclean et al. 2007; Echodu et al. 2015).

Nonetheless, the *T. congolense* population structure may be affected by various factors, including host species and geography. In fact, host selection certainly affects the population structure of *T. brucei* (MacLeod et al. 2000). Simo et al. (2014) reported that the *T. brucei* population circulating in tsetse flies and domestic mammals of the Fontem district of Cameroon was subdivided according to host species. Smetko et al. (2015) further noted that AAT is a potential driver of selection in West African cattle as a result of multiple crosses between trypanotolerant and susceptible breeds.

In terms of geography, different pathologies have been linked to distinct *T. brucei* *rhodesiense* strains circulating in the same district in Uganda (Maclean et al. 2007), one causing more severe symptoms associated with late-stage disease. Furthermore, the Ugandan and Southern-Kenyan populations of *T. brucei* have been shown to group into distinct genetic clusters according to location of collection, but retained strong evidence for recent gene flow between them (Echodu et al. 2015). Unlike in Cameroon (Simo et al. 2014), collected isolates did not cluster by host species, date of collection, or parasite subspecies.

As there is ample evidence showing that the population structure of *T. brucei* is dependent on a multiplicity of factors, the same may apply to *T. congolense* and be reflected on the VSG repertoire. In this dataset, whilst there are examples of strains with shared geography, host, and collection time that are genetically close and show similar VAPs (e.g. samples from Kabete isolated in 1976), there are also strains from the same region that do not (e.g. samples IL3954 and IL2281, which were collected from the same region and species on different years, but have closely related VAPs, even though they are not genetically close). These results show that the population genetics structure of *T. congolense* is complex and thus requires extensive, systematic sampling to be well understood. Yet, they suggest that the VSG repertoire is under independent evolutionary pressures and might evolve at a different rate to the entire genome, perhaps through recombination of closely related VSGs. This brings exciting research opportunities regarding the basis of recombination constraints in *T. congolense* compared to *T. brucei* and how VSG sequences are evolving at a genomic scale, including the importance of arbitrary mutagenic events and intra-phylotype homologous recombination in *T. congolense*.

2.4.3 The VAP methodology

The ability to profile antigenic diverse genes is an objective for researchers of all organisms expressing antigenic variation. The complexity, conservation and size of the gene repertoire are important aspects to note when developing a tool to quantify antigenic diversity.

In *P. falciparum*, profiling of *var* genes was achieved through a population genomic framework (Barry et al. 2007). Unlike the VAP, the *var* gene population genomic framework does not assess variation throughout the full gene, but rather uses only the ubiquitous 500- bp fragment marker, the Duffy binding like alpha (DBL α) domain, to sample the diversity in natural *P. falciparum* populations. This method lacks the ability to explain the global range of *var* gene diversity since it would require too intense sampling, but the issue of partial repertoire sampling is overcome by computing a cumulative diversity curve and thus estimating the number of isolates required for an accurate representation of diversity. Although this approach fails to include all the global diversity of *var* genes, due to the high degree of variation in the population, it has been helpful in linking *var* gene diversity to geography and uncovering small scale spatial diversity, represented by country-

specific repertoires, with little overlap among them (Chen et al. 2011). By contrast, the VAP can not only handle the global diversity of *T. congolense*, but has also shown that the *T. congolense* population VSG repertoire is stable and not strongly geographically-defined, in this sample set at least.

For profiling, I have used protein motifs to differentiate between phylotypes, but the use of protein groups to classify variant repertoires is not unprecedented. Immunoproteogenomics is an area of growing interest as it combines NGS and mass spectrometry (MS) to predict antibodies interacting with specific antigens. Accurate MS searches require accurate databases, but predicting antibody databases from NGS is not easy due to both antibody sequence diversification by recombination and mutations and NGS read error rate. The IgRepertoireConstructor corrects errors in immunosequencing reads to construct an immunoglobulin repertoire validated by MS. In the study, rather than single antibodies, antibody clones (or groups) are quantified so that better correlations between genomics-based and proteomics-based quantifications can be achieved. This tool has solved the problem of analysing the highly variable and repetitive antibody subfamilies and has moved the field towards studies of evolution of antibody repertoires (Safonova et al. 2015). Likewise, the VAP has solved the issues of handling the large numbers of VSGs by conventional population genomics methods, by presenting a faster, more sensitive and precise alternative. I anticipate the VAP to be useful in the association between disease phenotypes and host range with particular VSGs; tracking and characterising VSG expression in natural longitudinal and experimental infections; and in any antigen variation-associated study, such as non-variant VSG-associated genes knockouts.

Being able to profile antigens in a fast and high-throughput manner from NGS data with little need for bioinformatic processing is a key requirement of the VAP. I have shown that the VAP is effective by comparing its performance to manual, BLAST-based curation. For benchmarking, VAPs were obtained manually by sequence similarity search and compared to the motif-based VAP results. Interestingly, higher total numbers of VSGs were recovered using the latter than by the former, due to the high length thresholds applied to the sequence similarity search (i.e. 150 amino acids). These constraints were essential to account for sequence duplication and partial sequences. The automated process, however, did not require such considerations because it relies on fully probabilistic profile hidden Markov models (HMM), which are specific non-ambiguous structural models for homology searches

(Eddy 2004). The software used, HMMER3 (Finn et al. 2011), has the ability to accept multiple domains and sequence fragments, and allows insertions and deletions anywhere in the sequence. In contrast, BLAST can be deceived in the analysis of proteins with multiple distinct domains and high sequence variability (Eddy 1998). Homology searching through BLAST is the most widely used tool for sequence characterisation (Pearson 2013). However, for variable gene families its limitations increase considerably. This analysis revealed issues with the sequence similarity search method and showed that the VAP is faster, more sensitive and precise, and avoids many bioinformatic processing intersteps. The VAP method is better at first, distinguishing between partial sequences belonging to the same gene, and second, grouping together sequences belonging to phylotypes containing distantly-related sequences, but sharing a CTD (i.e. phylotypes 4, 8, 9).

Currently, the VAP deals with assembled contigs, but, ideally, it would be applied directly to raw DNA sequencing reads. However, motif discovery in raw reads would be influenced by sequencing coverage at that particular region. Although overall sequencing coverage can be easily calculated and accounted for in the VAP algorithm, this would assume equal sequencing depth across the genome, which cannot be verified, especially in repetitive regions such as the subtelomeres and telomeres where VSGs are mainly harboured. This would result in regions with higher sequencing coverage being preferred in the profile. Performing a *de novo* assembly provides unique contigs that can be searched for the motifs and accurately quantified. This computerised process avoids the need for mapping and phylogenetic analysis of individual samples. Nonetheless, as research evolves, the VAP methodology can be improved to incorporate contigs from long-read sequencing; through the segmentation of the current phylotypes into smaller, more specific ones, improving the sensitivity of the VAP; and through the attribution of functional and/or epidemiological meaning to each phylotype. In the future, I envisage particular VSG signatures, or VAPs, to be associated with specific disease phenotypes, disease outcomes and host use. As these relationships are unravelled, the VAP becomes more and more relevant in the understanding of host-parasite interactions.

2.4.4 Conclusion

This chapter showed that the *T. congolense* VSG repertoire could be dissected into 15 phylotypes, which accommodate the species variation observed to date. Given the ancestry of the phylotypes and the strong purifying selection on them, I believe they will accommodate any future *T. congolense* strain also. These phylotypes were used to develop variant antigen profiling, a bioinformatic approach to quantify antigenic diversity in *T. congolense* from high-throughput DNA sequencing data. VSGs have long been described as intimately linked to virulence and pathology, but they are highly dynamic and refractory to large-scale analysis. This study has revealed the extent of global antigenic diversity in *T. congolense* and provided the first approach to its high-throughput analysis in any population or experimental setting. Comparison of VAPs across *T. congolense* isolates showed that, although phylotype proportions are broadly conserved, there is individual variation in the repertoire, such that the VSG profile might become an epidemiological marker. Yet, three questions arise from the work presented in this chapter: first, whether the VAP can be applied to functional studies of VSG expression; second, whether all 15 phylotypes indeed encode functional VSGs; and third whether the VAP can be applied to other African trypanosome species. These points will be discussed over the next three chapters.

Chapter 3. The metacyclic VSG repertoire of *Trypanosoma congolense* 'savannah' Tc1/148

3.1 Introduction

After ingestion of an infected blood meal by the tsetse fly, *T. congolense* differentiates into procyclic form, colonising the midgut and rapidly proliferating. During BSF-PCF differentiation, the parasite coat is completely modified with the replacement of VSG with procyclin. After a successful midgut infection, *T. congolense* parasites migrate anteriorly towards the mouthparts, crossing the proventriculus valve, in a journey that can take 13-53 days, depending on parasite strain (Dale, 1995) and fly permissiveness (Haines, 2013). *T. congolense* infections have higher transmission rates than *T. brucei* (93 % vs. 25 %) (Peacock, 2012) due to morphological and physiological constraints: only short PCF can cross the proventriculus to the salivary glands and there is higher attrition in crossing the hypopharynx towards the salivary glands than to settle in the proboscis (Peacock, 2012).

In the mouthparts, PCFs differentiate into epimastigotes, following cell division, re-positioning of the kinetoplastid and attachment to the epithelial layer (Peacock, 2012). Differentiation into epimastigotes is also different between *T. brucei* and *T. congolense*. In *T. brucei*, cell division is asymmetrical, producing one short ($13 \pm 1 \mu\text{m}$) and one long epimastigote ($42 \pm 2 \mu\text{m}$) of moderately constant lengths (Van Den Abbeele et al. 1999). In *T. congolense*, asymmetrical division is not clear and the length variability of the daughter cells is high, which occasionally gives rise to extremely long epimastigotes (Peacock et al. 2012). Eventually, epimastigotes differentiate into infective metacyclics, detaching from the mouthparts and becoming small and very motile. Although the details of the epimastigote-metacyclic differentiation are not clear, parasites become considerably smaller (from 30 to $13\mu\text{m}$), arrest most gene transcription and start expressing VSG known as mVSG. *T. congolense* metacyclics primarily colonise the hypopharynx and sometimes the labrum (Gibson et al. 2017), so that they can be injected with the saliva into the mammal host during the fly bloodmeal. Although mVSG expression is essential for survival in the mammal host in the early stages of infection (Barry & McCulloch

2001), not much is known about either the surface coat composition of the metacyclics or the mechanisms of mVSG expression.

In 1978, Le Ray et al. showed that *T. brucei* metacyclic populations expressed multiple VSGs, lasting for up to 5 days after infection in the mammal host. These findings were corroborated by further studies with monoclonal antibodies by Hajduk et al. (1981). Shortly after, Esser et al. (1982) showed that metacyclic parasite populations from different flies expressed a shared set of antigen types even though they were infected with BSF expressing distinct antigens. Together, these data suggested that whilst *T. brucei* metacyclic populations expressed multiple VSGs, these were specific and potentially limited. This raised new questions, such as whether these restricted repertoires were conserved across time, strains, and cyclical transmissions. The analysis of the metacyclic antigens of parasites from a 20-year period in East Africa showed that *T. brucei* populations change their mVSG repertoire progressively over time, although the lost variants can still be expressed as bloodstream antigens in the early stages of infection (Barry et al. 1983). The finding that mVSG expression in *T. brucei*, unlike in the bloodstream, was programmed was corroborated by a study showing that a restricted set of up to 27 antigen types was repeatedly expressed regardless of the antigenic type expressed by the parasites ingested by the fly (Turner et al. 1988). These observations opened up new lines of research, for example, to understand how *T. brucei* could repeatedly select such a small set of VSGs in its infective stage; or how they could avoid the emergence of gradual immunity from the mammal host.

An attempt to answer the first question raised the hypothesis that mVSGs were expressed from a single expression site, which was supported by (1) the absence of gene rearrangement during mVSG expression, (2) very few sequence rearrangements around the MES that could move the mVSG to alternative expression sites, (3) the very short (and sometimes absent) 70 bp repeats upstream of mVSG, which are known to facilitate VSG conversion, and (4) mVSGs and bloodstream VSGs being independently controlled (Lenardo et al. 1986). Tetley et al. (1987) further showed that the heterogeneous mVSG population is present *ab initio* as activation occurs randomly after trypanosome division. Additionally, this confirmed that different mVSGs are expressed simultaneously in the population, which contrasts to the sequential activation in the bloodstream stage.

Another striking difference is the activation mode. Whilst bloodstream VSG are switched mostly by gene duplication events, mVSGs are activated *in situ* (Graham et al. 1990). Unlike the BES, the MES has few or no 70 bp repeats and has low homology with other telomeres (Graham et al. 1990). The 70 bp repeats are the anchor points for VSG transposition between the subtelomeres and the expression site, contributing to VSG switching. Similarly, lower sequence homology between the metacyclic and the bloodstream expression sites contributes to a lower frequency of homologous recombination and non-specific telomeric exchange, which increases stability in the MES and promotes the conservation of a common mVSG repertoire (Graham et al. 1990). The MES is also exclusively transcriptionally and developmentally regulated, being transcribed as a short monocistronic unit of 13 to 15 kb (Graham & Barry 1995). So far, the mVSG is the only gene transcribed monocistronically in African trypanosomes (Graham & Barry 1995). Whilst the BES is rich in ESAGs, the MES is not, being composed of a 426 bp promoter, a 3 kb gap containing the transcription initiation site, a pyrimidine-rich region and a short 70 bp repeat that may play a role in duplicating the mVSG to the BES (Graham & Barry 1995). The promoter itself is more complex than that of the BES, but contains an essential element that, on its own, can act as a BES promoter (Graham et al. 1998). Furthermore, it is recognised by the same class I transcription factor A (CITFA) required for BES transcription (Kolev et al. 2017). Its consensus sequence has been described both *in vivo* and *in vitro* by Ginger et al. (2002).

In *T. congolense*, the mVSG repertoire in a parasite population is also thought to be composed of multiple VSGs and limited to 12 antigen types (Crowe et al. 1983). Using mouse monoclonal antibodies raised against metacyclic parasites from the West African strain TREU1290, metacyclic parasite populations were labelled by indirect immunofluorescence, which showed that 12 antibodies were sufficient to label the entire population. Furthermore, the antigen types repertoire remained constant in composition *in vitro* for 6 weeks, although the relative frequencies of individual antigen types were variable (Crowe et al. 1983). Similarly to what was shown for *T. brucei*, the metacyclic antigen types were the first to appear as bloodstream antigens in *T. congolense*, but they seemed to be more stable, remaining exclusive in the bloodstream for nine days (Crowe et al. 1983; Luckins et al. 1994). Yet, the repertoire might be variable between isolates as none of the antibodies used to label the metacyclic populations of TREU1290 worked on East African strains (Crowe et al. 1983).

Based on antibody labelling, these studies lack the ability to distinguish genes belonging to the same antigen type, so it is possible that the 12 different VATs include multiple genes. In 1992, Eshita et al. cloned and sequenced 2 mVSGs, showing that they were 428 and 447 amino acids in length and homologous to bloodstream VSGs. The analysis of their genomic context revealed they were in a conserved region of at least 27 kb associated with the telomeres. Furthermore, the first was shown to be activated without gene rearrangement, whilst the second showed evidence for genomic alteration between metacyclic and bloodstream conversion in at least one copy (Eshita et al. 1992). These results suggest that the composition of the *T. congolense* mVSG repertoire and the mechanisms of *T. congolense* mVSG expression and activation remain largely unknown and may differ from what has been established in *T. brucei*. Substantial differences in life cycles in the tsetse fly, the VSG genomic repertoire and its evolution further corroborate the potential differences in regulation and/or expression of mVSGs.

As introduced in Chapter 2, the *T. congolense* genomic VSG repertoire is divided into 15 phylotypes whose relative proportions vary in the population. The VAP can rapidly quantify these proportions in the genome and put any given repertoire in context of the strains sampled to date. This approach can be of great use for population genomic studies, but can potentially be modified for the analysis of transcriptomic data. By incorporating transcript abundance values, transcriptomic VAPs can be a measure of VSG expression and thus become useful for various functional experiments. This chapter aims to characterise the expressed mVSG repertoire of a fly transmissible strain of *T. congolense* (Tc1/148) by applying the variant antigen-profiling tool. This approach will provide a realistic example of how the VAP can be useful in functional experiments of VSG expression and target a long-standing question on the characteristics of the mVSG repertoire in *T. congolense*.

The specific aims are:

1. To extend the VAP to the analysis of transcriptomic data to allow the quantification and characterisation of VSG expression.
2. To produce transcriptomic and proteomic VAPs for metacyclic populations of Tc1/148 obtained from experimental fly infections to characterise the mVSG repertoire of metacyclic *T. congolense* parasites.
3. To investigate the variability of the mVSG repertoire of parasites from vector biological replicates and compare it what is established for *T. brucei*.

3.2 Methods

3.2.1 *T. congolense* Tc1/148 strain

T. congolense 'savannah' Tc1/148 (MBOI/NG/60/1-148) (Young & Godfrey 1983) mouse blood stabilates were obtained from the Department of Vector Biology and Department of Parasitology of the Liverpool School of Tropical Medicine, UK. This Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accession NHOR01000000.

3.2.2 Tsetse fly infection and rearing

Pooled flies:

Experimental teneral male tsetse flies (*G. morsitans morsitans*) were infected 12-48 hour post-eclosion with 5×10^5 procyclic forms per ml⁻¹ in sterile defibrinated horse blood supplemented with 10 mM glutathione, via a silicone membrane by pouring 4mls of blood onto three distinct spots on a feeding tray covered with a feeding membrane (Moloo 1971). Flies were allowed to feed for 11 min at 27 °C and in the dark. One day after infection, unfed flies were sorted and removed. Remaining flies were maintained at 26 °C (+/- 1 °C) and 65-75 % relative humidity and fed every 2 to 3 days with normal sterile defibrinated horse blood. Flies were killed by decapitation at day 28 post-infection.

Individual flies:

A frozen stabilate of 1ml of infected mouse blood (20 % glycerol and parasitaemia of 10^3 parasites/ml) was thawed and mixed with defibrinated horse blood at a 1:10 ratio. The sample was used to infect 100-180 teneral male tsetse flies (*G. m. morsitans*) 12-48h post-eclosion following the procedure described in the previous section. Flies were allowed to feed for 11 min at 27 °C and in the dark. One day after infection, unfed flies were sorted and removed. Remaining flies were fed every 2-3 days with sterile defibrinated horse blood. Flies were killed by decapitation at day 28 post-infection. This procedure was repeated twice for the metacyclic enrichment (see section 3.2.4).

3.2.3 Fly dissection

Pooled and individual flies for direct RNA extraction

In the first two experiments, RNA was extracted directly from the dissected mouthparts. Flies were dissected 28 days after infection and 3 days of starvation. Midguts were dissected in a drop of PBS and viewed under the light microscope (LM) (100x magnification). Mouthparts from flies with midgut infections were dissected into a separate drop of PBS and visualised under the light microscope (10X magnification). Dissections were performed according to an adaptation of the description of Peel (1962). Flies were chilled to 4 °C to arrest their movement and kept on ice until dissection. Using very sharp, anti-magnetic, straight forceps (Sigma-Aldrich, UK), flies were decapitated and the proboscis detached by holding it strongly and pulling between the head and its base. The proboscis was placed in a drop of PBS, opened to separate the labrum, the hypopharynx and the labium, and visualised under the light microscope (10X magnification). When metacyclic parasites were visible, the hypopharynx and the parasite suspension that was released during dissection were collected and frozen in liquid nitrogen.

Pooled flies for metacyclic enrichment

In the third experiment, RNA and protein was extracted from metacyclic-enriched parasite populations collected from tsetse mouthparts. Flies were dissected 29 days after infection and 3 days starvation. Mouthparts were dissected into a drop of PBS (method 1, see section 3.2.3) or glucose separation buffer (44 mM NaCl, 57 mM Na₂HPO₄, 3 mM KH₂PO₄, 55 mM glucose, pH 8.0 at 20 °C) (method 2, see section 3.2.3) and visualised under the light microscope (10X magnification). Dissections were performed according to the description of Peel (1962) and as described in the previous section. When metacyclic parasites were visible, the hypopharynx was broken down to release the parasites and the parasite suspension collected by aspiration and kept on ice until the metacyclic enrichment procedure.

3.2.4 Metacyclic enrichment from fly mouthparts

Metacyclic parasites were separated by two methods: the first using poly-L Lysine microscope slides to which epimastigote parasites should preferentially adhere to, and the second using anion exchange chromatography in an adaptation of the previously described DE52 cellulose separation method (Lanham & Godfrey 1970).

Method 1 - Poly-L lysine adhesion

Parasite suspensions in PBS were separated into nine 1.5 ml tubes (Eppendorf, UK), and kept on ice. The parasite suspensions were spread onto a poly-lysine microscope slide (Sigma-Aldrich, UK) and kept on a humidity chamber to allow for epimastigote adhesion. After an incubation of 10 min, the solution left on the slide was carefully recovered with a pipette. The resulting parasite solutions from the 9 poly-L lysine slides were pooled into a 1.5 ml tube and frozen in liquid nitrogen. The epimastigote-metacyclic ratio before and after enrichment was calculated manually by cell counting on a haemocytometer.

Method 2 - DE52 cellulose column separation

Pre-swollen DE52 cellulose granules were activated previous to use through the following steps: cellulose was mixed with 4 volumes of 0.5 M HCl in a beaker and allowed to stand for 30 min with brief stirring every 10 min. After the last settle, most of the supernatant was removed by aspiration and the cellulose transferred to a glass funnel lined with filter paper. Cellulose was washed with distilled water by gravity until the effluent was pH4 (measured with pH strips with ranges of 1-4 and 4-6). Cellulose was returned to the beaker and mixed with 4 volumes of 0.5 M NaOH in a beaker and allowed to stand for 30 min with brief stirring every 10 min. After the last settle, most of the supernatant was removed by aspiration and the cellulose transferred to a glass funnel lined with filter paper. Cellulose was washed with distilled water by gravity until the effluent was pH8 (measured with pH strips with range of 7-10). Cellulose was transferred to a clean beaker containing 1 volume of separation buffer (44 mM NaCl, 57 mM Na₂HPO₄, 3 mM KH₂PO₄, pH 8.0 at 20 °C). Using pH strips, pH was titrated to 8.0 with orthophosphoric acid. Cellulose was transferred to a storage bottle with an equal volume of separation buffer, added 0.001 volumes of chloroform as a preservative and stored at 4 °C.

Parasites were kept on ice during dissections and during column preparation. To prepare the columns, 200 µl DE52-cellulose was added to a Poly-Prep® Chromatography Column (Bio Rad, UK) and allowed to settle. Columns were equilibrated with 400µl glucose separation buffer (44 mM NaCl, 57 mM Na₂HPO₄, 3 mM KH₂PO₄, 55 mM glucose, pH 8.0 at 20 °C) by gravity to remove chloroform. 50 µl of trypanosome mixture was added to the column and the eluate recovered on ice. The column was washed with 5 volumes of glucose separation buffer and eluate recovered on ice. The eluate was checked by microscopy and metacyclic:

epimastigotes ratio calculated. Parasites were snap frozen on liquid nitrogen and kept at -80 °C until RNA and protein extractions. The epimastigote-metacyclic ratio before and after enrichment was calculated manually by cell counting on a haemocytometer.

3.2.5 RNA and protein extraction

Samples from Pooled, Individual flies, and metacyclic-enriched parasite solutions from Method 1

Total RNA and protein from dissected hypopharynxes were extracted with the AllPrep™ RNA/Protein Kit (Qiagen, UK) according to the manufacturer's protocol, yielding 48-213 ng of total RNA per sample. RNA was quantified with Qubit® fluorometric RNA quantitation (RNA HS Assay Kit) (Life Technologies, UK) and purity analysed using the NanoDrop™ spectrophotometer (ThermoFisher, UK).

Pooled samples after metacyclic enrichment

Total RNA from parasite suspensions was extracted with the RNaseasy Kit (Qiagen, UK), yielding a total RNA output between 48 and 246 ng. RNA was quantified with Qubit® fluorometric RNA quantitation (RNA HS Assay Kit) (Life Technologies, UK) and purity analysed using the NanoDrop™ spectrophotometer (ThermoFisher, UK).

3.2.6 RNA sequencing

To proceed with RNA sequencing, RNASeq libraries were prepared by the Centre of Genomic Research (Liverpool, UK) using the NEBNext® Ultra™ II Directional RNA Library Prep Kit with poly-A selection from total RNA (Poly-(A) mRNA Magnetic Isolation Module) (New England Biolabs, UK). The protocol was followed as per manufacturer's instructions and included mRNA isolation using AMPure XP Beads (Agencourt, UK), bidirectional cDNA synthesis and purification, cDNA end repair and adaptor ligation, and PCR enrichment of adaptor ligated cDNA. Following preparation, library quality was assessed on the Bioanalyzer® (Agilent High Sensitivity Chip, Agilent Technologies, USA).

For each infection, RNASeq libraries were sequenced on a single lane of the HiSeq2500 platform (Illumina Inc, USA) as 150 paired ends, producing 280 million mappable reads, at the Centre of Genomic Research (Liverpool, UK).

3.2.7 Sample preparation for proteomics

Pooled and individual samples

To solubilise protein, tsetse mouthpart material was sonicated in 50 mM ammonium bicarbonate on ice before centrifugation at 12,000 x g for 10 min to pellet debris. Proteins present in the supernatant were precipitated with 5 volumes of ice-cold acetone and incubated overnight at -20 °C. Samples were then centrifuged at 12,000 x g for 10 min to pellet protein.

Protein was re-solubilised in 0.1 % (w/v) Rapigest (Waters) in 50 mM ammonium bicarbonate then incubated at 80 °C for 10 min before reduction with 3 mM dithiothreitol (DTT) at 60°C for 10 min and alkylation with 9 mM iodoacetimide at room temperature for 30 min in the dark. Proteomic-grade trypsin (Sigma-Aldrich) was added at a protein-trypsin ratio of 50:1, and samples were incubated at 37 °C overnight. TFA was added to a final concentration of 0.5 % (v/v) to remove Rapigest. Peptide samples were centrifuged at 12,000 x g for 30 min to remove precipitated material.

Pooled samples after metacyclic enrichment

To solubilise protein, tsetse mouthpart material was sonicated in 50 mM ammonium bicarbonate on ice before centrifugation at 12000 x g for 10 min to pellet debris. Proteins present in the supernatant were precipitated with trichloroacetic acid to precipitate any soluble proteins that may have lysed during the sample freezing. The resulting protein fractions (pellet and supernatant) were pooled and treated with PNGase F for de-glycosylation, with a fetuin positive control. The samples were split and treated with either trypsin or chymotrypsin as described in the previous section.

Washed pooled samples exclusively for proteomics

To reduce the tsetse background obtained in the MS analysis, the parasite suspensions were washed multiple times by centrifugation. 106 infected-mouthparts were dissected and washed in PBS to release parasites. Parasite suspension was washed in 100 µl of PBS by sequential centrifugation at 1700 g, for 5 min at 4 °C

followed by a fast spin at 5,200 g for 30 seconds, until no parasites were visible in the eluate by light microscopy. The parasite pellet was snap frozen in liquid nitrogen and stored at -80 °C until MS analysis.

GPI-PLC treatment

Endogenous GPI-PLC cleaves the GPI-anchor, resulting in the de-attachment of GPI-anchored proteins, such as the VSG, from the parasite membrane. The procedure, described previously in *T. brucei* by Sunter et al. (2013), was carried out to increase the specificity of the MS analysis towards VSG.

37 infected-mouthparts were dissected and washed in PBS to release parasites. Parasite suspension was washed in 100 µl of PBS by sequential centrifugation at 1,700 g, for 5 min at 4 °C followed by a fast spin at 5,200 g for 30 seconds, until no parasites were visible in the eluate by light microscopy. The pellet was resuspended in 200 µl of 1 mM TrisHCl (pH8.0) and SigmaFast Protease Inhibitor cocktail (catalogue number: S8820) (Sigma, UK), kept at 0 °C for 5 min, and then incubated in a water bath at 37 °C for 20 min. The suspension was centrifuged at 4 °C for 3 min at maximum speed and the pellet snap frozen in liquid nitrogen and stored at -80 °C until MS analysis.

3.2.8 NanoLC MS ESI MS/MS analysis

Peptides were analysed by on-line nanoflow LC using the Ultimate 3000 nano system (Dionex/Thermo Fisher Scientific). Samples were loaded onto a trap column (Acclaim PepMap 100, 2 cm × 75 µm inner diameter, C18, 3µm, 100 Å) at 5 µl.min⁻¹ with an aqueous solution containing 0.1 % (v/v) trifluoroacetic acid and 2 % (v/v) acetonitrile. After 3 min, the trap column was set in-line an analytical column (Easy-Spray PepMap® RSLC 50 cm × 75 µm inner diameter, C18, 2 µm, 100 Å) fused to a silica nano-electrospray emitter (Dionex). The column was operated at a constant temperature of 35 °C and the LC system coupled to a Q-Exactive mass spectrometer (Thermo Fisher Scientific). Chromatography was performed with a buffer system consisting of 0.1 % formic acid (buffer A) and 80 % acetonitrile in 0.1 % formic acid (buffer B). The peptides were separated by a linear gradient of 3.8 – 50 % buffer B over 90 min at a flow rate of 300 nl/min. The Q-Exactive was operated in data-dependent mode with survey scans acquired at a resolution of 70,000 at m/z

200. Up to the top 10 most abundant isotope patterns with charge states +2 to +5 from the survey scan were selected with an isolation window of 2.0 Th and fragmented by higher energy collisional dissociation with normalized collision energies of 30. The maximum ion injection times for the survey scan and the MS/MS scans were 250 and 100 ms, respectively, and the ion target value was set to 1E6 for survey scans and 1E4 for the MS/MS scans. MS/MS events were acquired at a resolution of 17,000. Repetitive sequencing of peptides was minimized through dynamic exclusion of the sequenced peptides for 20 s.

3.2.9 Protein identification and quantification

Thermo RAW files were imported into PEAKS studio 7 software [Bioinformatics Solutions Inc., Waterloo, ON, Canada, (Ma et al. 2003)]. Peaks were picked by the software using default settings. Tandem MS data were searched against translated ORFs from *T. congolense* Tc148 (12,174 sequences). The search parameters were as follows: precursor mass tolerance was set to 10 ppm and fragment mass tolerance was set as 0.01 Da. Two missed tryptic cleavages were permitted. Carbamidomethylation (cysteine) was set as a fixed modification and oxidation (methionine) set as variable modification. Only proteins with score of >20 (-10 lgP) were considered significant matches. The false discovery rate was set at 1 %.

3.2.10 Transcriptome profiling

RNAseq reads were mapped to the tsetse fly genome (International Glossina Genome Initiative 2014) to deplete host reads using Bowtie2 (Langmead & Salzberg 2012) and the unmapped data was mapped to the *T. congolense* IL3000 genome. The bam file was used to estimate transcript abundance values using cufflinks (Trapnell et al. 2012). The VSG transcripts and their abundances were extracted from the cufflinks output, screened for the phylotypes motifs described previously, and transcriptomic VAPs were estimated by adjusting the relative frequency of each phylotype to the relative abundance of each transcript. To compare the stability of the composition of the VAPs to the background random variation and genomic compositions, the total pool of VSGs recovered in the study was used to create 24 randomised simulated VAPs based on 79 VSGs each.

3.2.11 Statistical analysis

The statistical comparisons between the BLAST and the VAP performances in recovering VSGs were done using the Pearson's correlation test. Outliers were identified using a threshold of 2σ with the function 'removeOutlier' in R (RStudio Team 2016). Outliers were manually inspected before removal. The statistical analysis of differential profile expression was done using Pearson's chi-squared test, F-tests were performed to analyse variance between transcriptomic and simulated data, and independent student t-tests were performed to detect statistical significances in phylotype relative abundances. All tests were performed in R (RStudio Team 2016).

3.3 Results

It is possible to culture all life stages of *T. congolense* using the reference strain IL3000, although cyclical infections still require infection of a mammal host due to the challenge in differentiating from metacyclic to bloodstream forms in culture. However, as the tsetse fly is highly refractory to trypanosome infection, the changes in physical environments and challenges during development in the vector have a critical impact on the parasite, likely affecting gene expression. Therefore, in this study a fly-transmissible strain of *T. congolense*, Tc1/148, was used and the parasites were recovered from experimentally-infected tsetse flies. All fly infections resulted in a mouthpart infection rate of 86 % or higher.

3.3.1 Trypanosome populations in the tsetse fly mouthparts

Parasites were visible in the mouthparts as early as day 10 post-infection. However, peak mouthpart colonization by metacyclic forms occurred at day 28 post-infection. The tsetse fly mouthparts were colonised by a mixture of epimastigotes and metacyclics in different stages of differentiation. Parasites of both life stages aggregated in patches throughout the hypopharynx, creating distinctly white areas inside the transparent hypopharynx (**Figure 21B**). In general, more than 75 % of the cells were epimastigotes, easily identified by their long size, attachment to the hypopharynx walls, and slow movement (**Figure 21C and 13D**). Yet, at the time of dissection, many epimastigotes were released to the outside environment. Metacyclic forms tended to colonise the internal part of the hypopharynx and a fine adjacent duct (**Figure 21C and D**). As previously described in the literature (Peacock et al. 2012), metacyclics were small and motile.

In fixed preparations, the observations made during dissections were confirmed. Epimastigotes were long and showed a posteriorly located kinetoplast very close to the nucleus (**EF1-2, Figure 22**). Some epimastigotes were much longer than usual, a feature which has been previously reported in prolonged infections (**EF3, Figure 22**) (Peacock et al. 2012). In contrast, metacyclics were 5-10 μm in length and the kinetoplast was positioned anteriorly and distantly from the nucleus (**MF1-2, Figure 22**).

An attempt to increase the metacyclic: epimastigote ratio in the mouthparts by delaying the day of dissection by 7 days (day 35 post-infection) resulted in a greater number of long epimastigotes (e.g. EF3, Figure 22) and not a visible difference in the metacyclic concentration. Therefore, subsequent dissections were performed at day 28 post-infection.

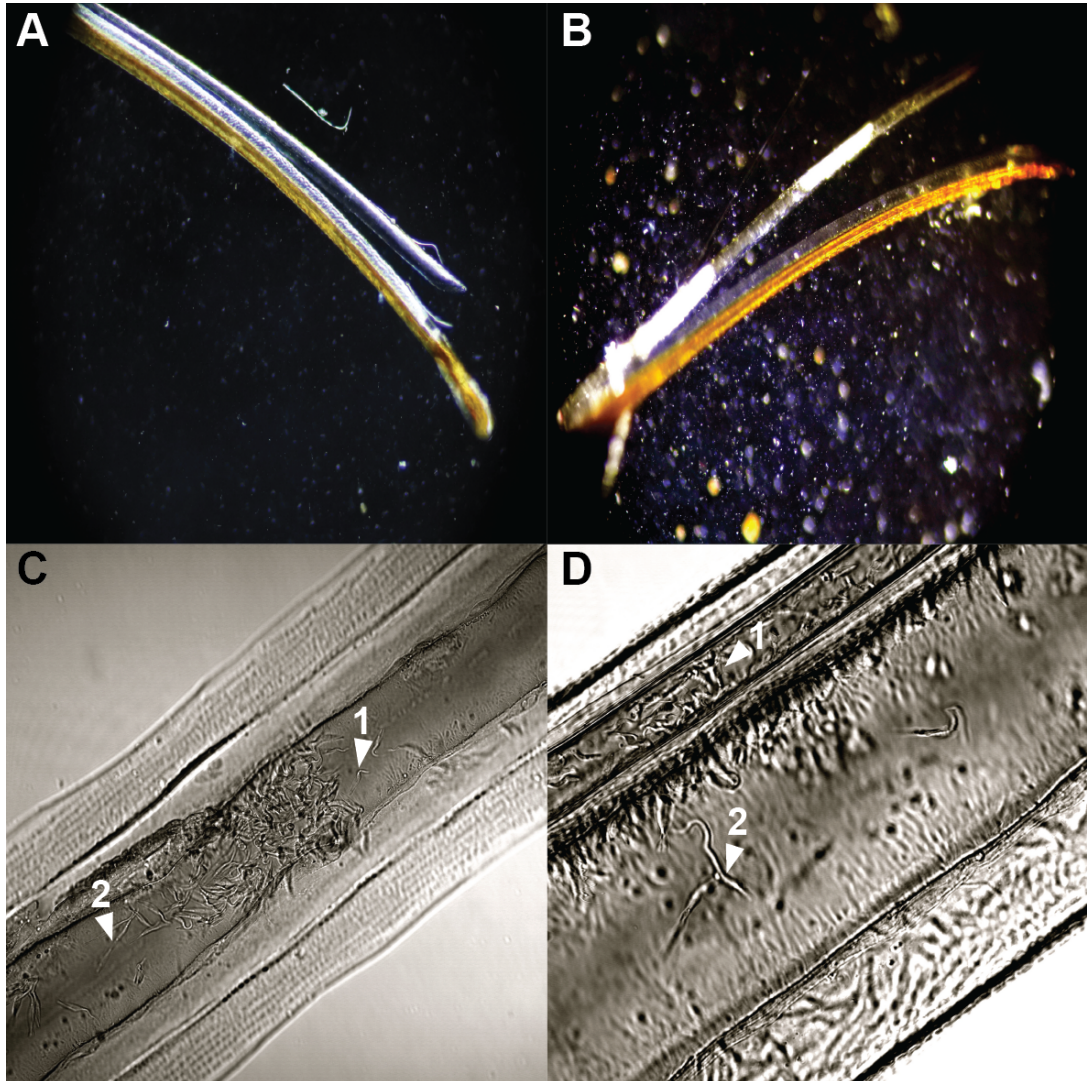


Figure 21 *T. congolense* in the tsetse fly hypopharynx. A. Uninfected tsetse mouthpart under light microscope (10 X). B. Infected tsetse mouthpart under light microscope (10X), showing two agglomerations of *T. congolense*. C. Live imaging of tsetse hypopharynx under the confocal microscope showing a mixed population of metacyclics (1) and epimastigotes (2). D. Live imaging of tsetse hypopharynx under the confocal microscope showing distinct localization of metacyclics in a fine duct above the hypopharynx (1) and epimastigotes in the hypopharynx (2).

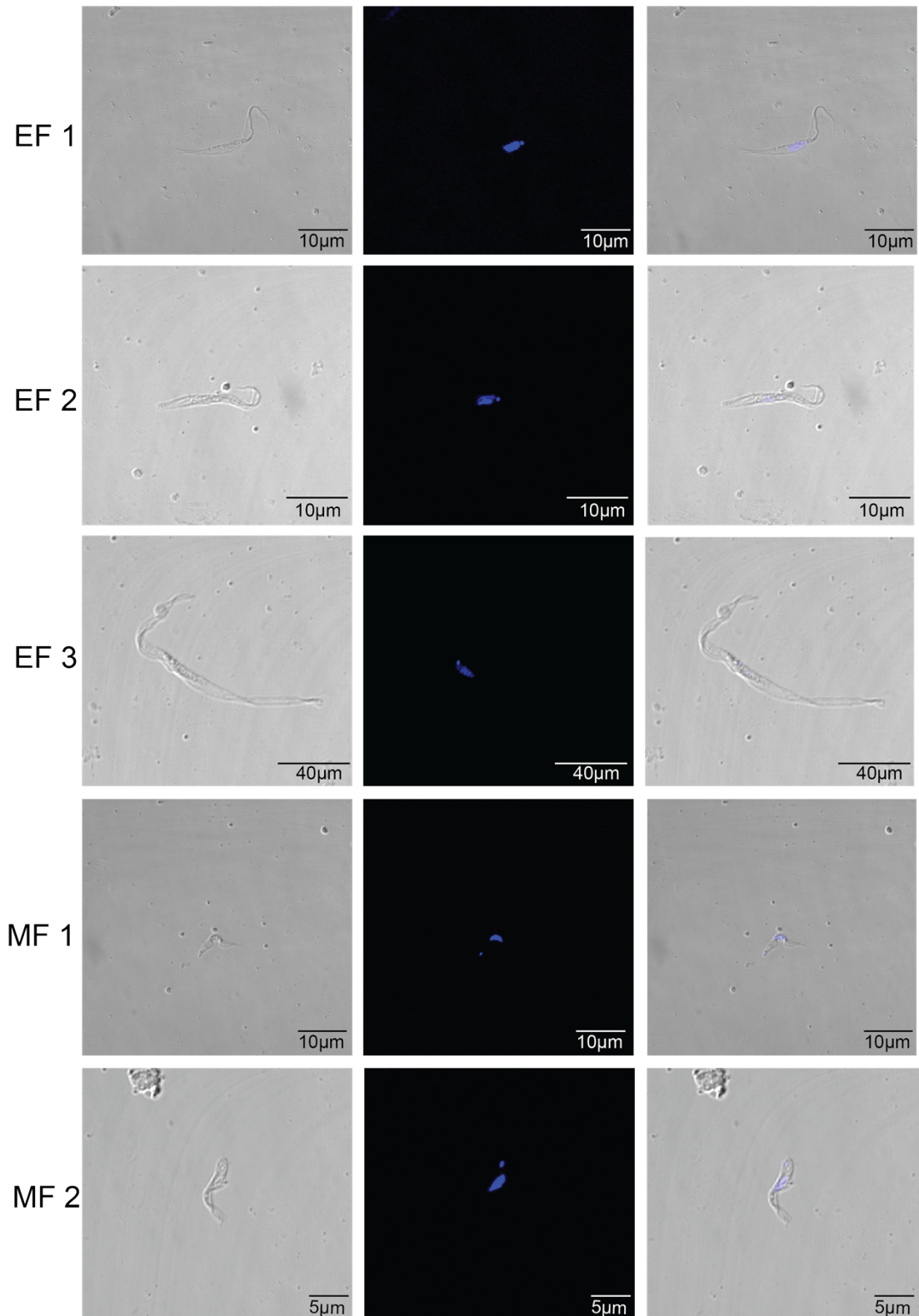


Figure 22 Immunofluorescence detection of trypanosome populations from the tsetse fly mouthparts on a confocal microscope. Trypanosomes were stained with DAPI (blue) to reveal nuclei, kinetoplast and their relative positioning. Populations are composed of regular epimastigotes (EF1-2), long epimastigotes (EF3) and infective metacyclics (MF1-2). Scale bars are shown in individual pictures.

3.3.2 Adapting variant antigen profiling to transcriptomic data

Adapting variant antigen profiling to transcriptomic data implies taking into account not only the relative frequency of each phylotype, i.e. how many transcripts of each phylotype are present in the transcriptome, but also the expression level of each transcript. This ensures that the VAP reflects VSG expression accurately for a given condition. To achieve this, the *de-novo* assembly used in the genomic VAP was replaced by transcript mapping to the strain genome and followed by transcript abundance estimation. Transcript mapping produces mapped RNAseq data in the form of a bam file that can be passed to software to quantify the expression level of protein coding genes. Here, Bowtie2 (Langmead & Salzberg 2012) and Cufflinks (Trapnell et al. 2012) were used in the first and second steps, respectively, resulting in a list of transcripts with the reference gene identifiers and the respective abundance values. The transcript identifiers are used to recover their nucleotide sequence, which is translated and screened for the phylotype motifs. Posteriorly, the relative frequency of each phylotype is adjusted for the abundance of each transcript, resulting in a weight-based VAP.

3.3.3 The expressed mVSG repertoires of trypanosomes

To understand the degree of variation in the VSG repertoire of different parasite populations and to evaluate whether successful transcriptomes and proteomes could be obtained from direct RNA and protein extraction of the tsetse fly hypopharynx, tsetse flies were infected with cultured procyclic forms and a transcriptome and proteome from 40 pooled mouthparts was produced. This first infection was done to establish if sufficient RNA could be recovered to produce a reliable VAP. We recovered 67 VSG transcripts, relating to various phylotypes, although the most abundant VSG transcript belonged to phylotype 8 (**Figure 23, infection 1**). Interestingly, mVSGs were not the most abundant transcripts in the samples. The fragments per kilobase of transcript per million mapped reads (FPKM) abundance of the 200 most abundant transcripts per sample ranged from 5,621.37 to 492.41. This list included 147 different functional annotations, of which ribosomal proteins, hypothetical proteins, histones, S-adenosylmethionine synthetase, cytochrome oxidase, elongation factors, alpha tubulin and nodulin-like proteins were the most frequent.

At this point, mass spectrometry analysis did not reveal any VSGs.

Although it was possible to produce a reliable transcriptomic VAP, the biological insights gained from pooled data are very limited because we cannot estimate the degree of variation between flies that might be caused by epigenetic effects. Therefore, transcriptomes and proteomes from twenty-four individual tsetse flies were produced. The second infection was done with a *T. congolense* strain Tc1/148 blood stabilate after one mouse passage, an effort to optimise infection rates and parasite colonisation. The transcriptomes contained 20.4-37.8 million reads per sample, of which 6 to 47 % mapped to the *T. congolense* Tc1/148 genome sequence. After transcript abundance estimation, the mapped reads resulted into 6462 to 11466 transcripts, of which 31 to 147 were VSGs (mean $\pm \sigma = 79 \pm 31$; FPKM = 103-634) (**Table 7**). As observed in the pooled transcriptomes, mVSGs were not the most abundant transcripts. The transcriptomic VAPs showed remarkably low variation among flies, but the VAP was distinct from the genomic profile (**Figure 23, infection 2**). All phylotypes were represented.

Table 7 Sequencing statistics and number and expression values of VSG transcripts recovered per transcriptome.

Sample ID	Read pairs	Transcripts	Maximum FPKM	VSG transcripts	Maximum FPKM
1	3.19E+07	10695	3906.16	110	127.93
2	2.87E+07	9920	2505.62	63	270.35
3	2.32E+07	11073	2114.95	108	125.93
4	3.79E+07	9554	1798.98	52	183
5	2.24E+07	9501	2981.22	52	152
6	2.14E+07	8524	1355.24	43	138.16
7	2.82E+07	9619	4104.68	75	156.65
8	2.91E+07	6493	5621.37	36	155.84
9	3.50E+07	8374	1952.73	62	159.18
10	2.52E+07	6462	3119.37	46	191.9
11	2.77E+07	9768	4817.29	54	128.46
12	2.04E+07	6628	4026.51	31	131.78
13	3.23E+07	10694	2093.85	120	102.88
14	3.36E+07	7335	2729.18	79	431.31
15	4.00E+07	11437	2012.88	147	176.31
16	3.79E+07	8952	2368.46	68	173.21
17	3.24E+07	10811	1869.47	114	213.47
18	3.29E+07	11466	3003.4	94	222.41
19	4.13E+07	10456	2213.25	90	186.71
20	2.67E+07	10490	1754.01	114	150.43
21	3.49E+07	10243	1455.93	105	131.23
22	3.31E+07	8734	3346.27	118	634.24
23	3.16E+07	7200	1608.64	53	181.42
24	2.35E+07	9415	2696.13	66	316.45

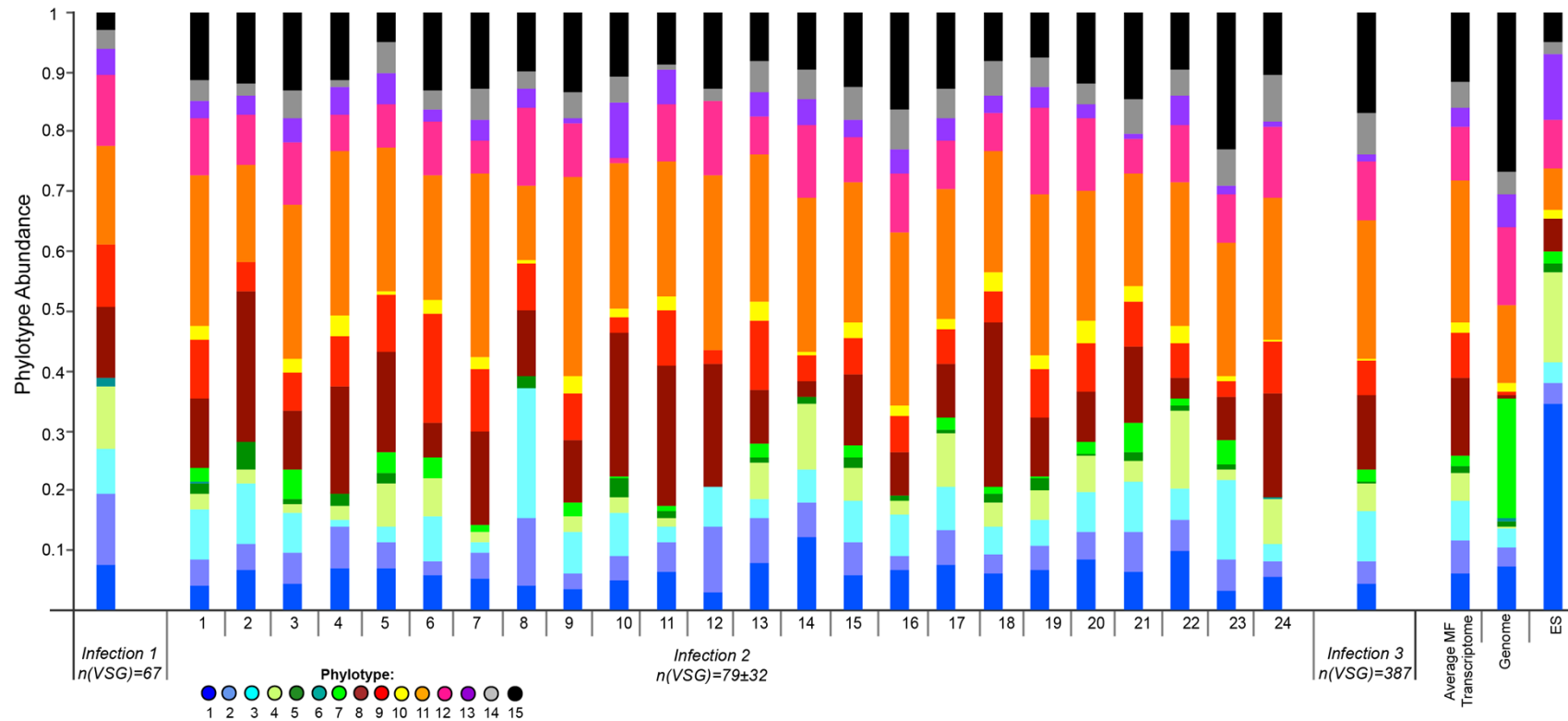


Figure 23 Transcriptomic Variant Antigen Profiles of trypanosomes extracted from tsetse mouthparts. VAPs from the transcriptomes are remarkably similar, yet significantly different from the genomic VSG repertoire (Poisson regression model, $p < 0.001$) and the VSGs found at the *T. congolense* telomeric expression sites. Infection 1 represents a sample of 40-pooled mouthparts; infection 2 represents 24 individual mouthparts; infection 3 represents a sample of 131-pooled mouthparts after metacyclic parasite enrichment by anion exchange chromatography. The genomic VAP represents the average profile of 24 sets of 79 VSGs randomly sampled from the genome of Tc1/148. Stacked columns are colour-coded by phylotype according to key. The number of VSG transcripts recovered in each sample infection is noted in the figure.

A set of simulated VAPs was created to investigate whether the transcriptomic VAPs from different flies were more consistent than expected by chance, and to determine under- and over-represented phylotypes in the transcriptomes compared to the genome. Using the full genomic VSG repertoire of Tc1/148, 24 sets of 79 VSGs were randomly selected and profiled to become comparable to the experimental data. The simulated data showed a high correlation with the genomic repertoire ($R^2 = 0.99$). After comparing the variance of each phylotype in the experimental transcriptomic and simulated genomic data, only phylotype 6 and 7 were less variable in the experimental data, suggesting that the metacyclic VSG profile is more variable than a random selection of genomic VSGs. However, as the metacyclic VAPs were significantly distinct from the genomic VAP (Poisson regression model, $p < 0.001$), the relative abundances of specific phylotypes were compared using two-sample, two-way, student t-tests. This analysis showed that seven phylotypes show significantly different relative proportions in the transcriptomes compared to the genome: phylotypes 7, 12, and 15 are under-represented in the transcriptomes (independent t-test, $p < 0.001$), whilst phylotypes 4, 8, 9 and 11 are significantly over-represented (independent t-test, $p < 0.001$) **(Figure 24)**.

In terms of magnitude, phylotypes 6, 7, 12 and 15 show a lower relative abundance in the transcriptomes; 6 and 7 have the most accentuated differences being under-represented by 97 and 91 %, respectively. Phylotypes 4, 8, and 9 are largely overrepresented in the transcriptomes by 1291, 2699, and 1305 %, respectively. Phylotype 11 shows an over-representation of 98 %. The abundance of the remaining phylotypes is not significantly different from their genomic extent. This indicates that those over-represented phylotypes may be preferentially expressed in the metacyclic stage.

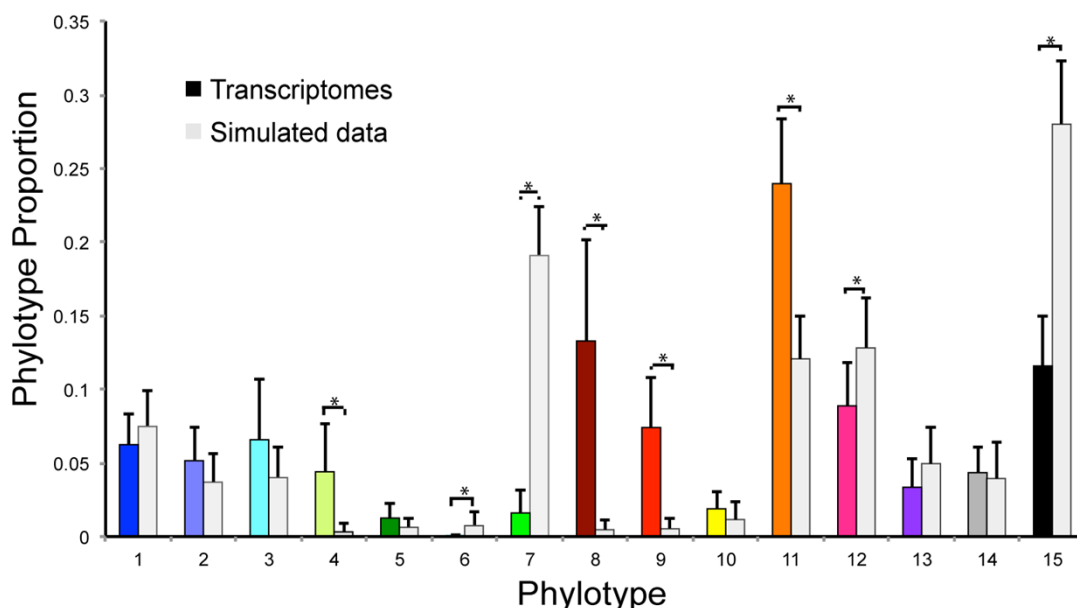


Figure 24 Comparison of average phylotype relative abundance (adjusted for transcript abundance) in transcriptomic samples and genomic profiles from a random selection of VSGs of Tc1/148 (mean \pm σ). Phylotype relative abundance in the transcriptomes is shown in bold; phylotype relative frequency in the random simulations from the genome is shaded lighter. Statistical analysis reveals that, in comparison to the genome, phylotypes 7, 12, and 15 are under-represented in the transcriptomes (independent t-test, $p < 0.001$), whilst phylotypes 4, 8, 9 and 11 are significantly over-represented (independent t-test, $p < 0.001$).

A closer analysis of phylotypes 11 and 8, which are the most abundant and the most over-represented in the metacyclic transcriptome, respectively, revealed important differences in composition. At the genome level, phylotype 11 contains 146 genes, of which 74 (50 %) were expressed across the sample set, with variable, yet generally low, transcript abundances (FPKM ranges from 8.38×10^{-5} to 69.84). Only one transcript (Tc14808730) is common to all samples, while 29 are sample-specific. This suggests that phylotype 11 is collectively abundant at the transcriptomic level due to multiple, randomly selected VSGs and not to any specific gene(s). In contrast, the relative abundance of phylotype 8 derives from two particular transcripts (98 % and 99 % identical to TcIL3000_0_09520 respectively), common to all samples and a third transcript common to 23/24 samples (99 % identical to TcIL3000_5_650). These correspond to 33.33 % of the phylotype's genomic repertoire (3/9) and have consistently high expression values, mostly within

the 6 most abundant VSG transcripts (**Figure 25**). Therefore, in contrast to phylotype 11, the profusion of phylotype 8 derives from reproducible expression of particular genes; their phylogenetic position is shown in **Figure 26**.

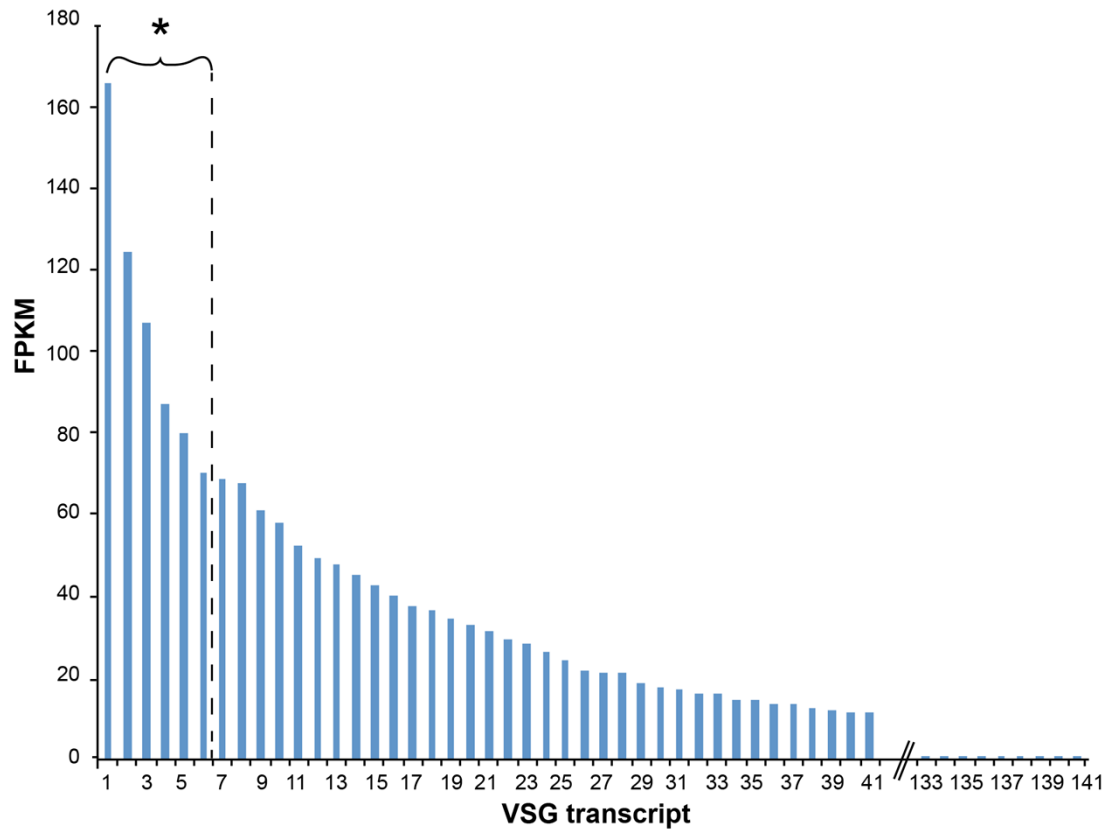


Figure 25 VSG transcript abundances. Bar graph shows the median FPKM values for the VSG transcripts recovered from the mouthpart transcriptomes (N = 24). VSGs are ordered by transcript abundance. Star indicates the presence of the majority of phylotype 8 members; 21 out of 24 samples have phylotype 8 members within the 6 most abundant transcripts.

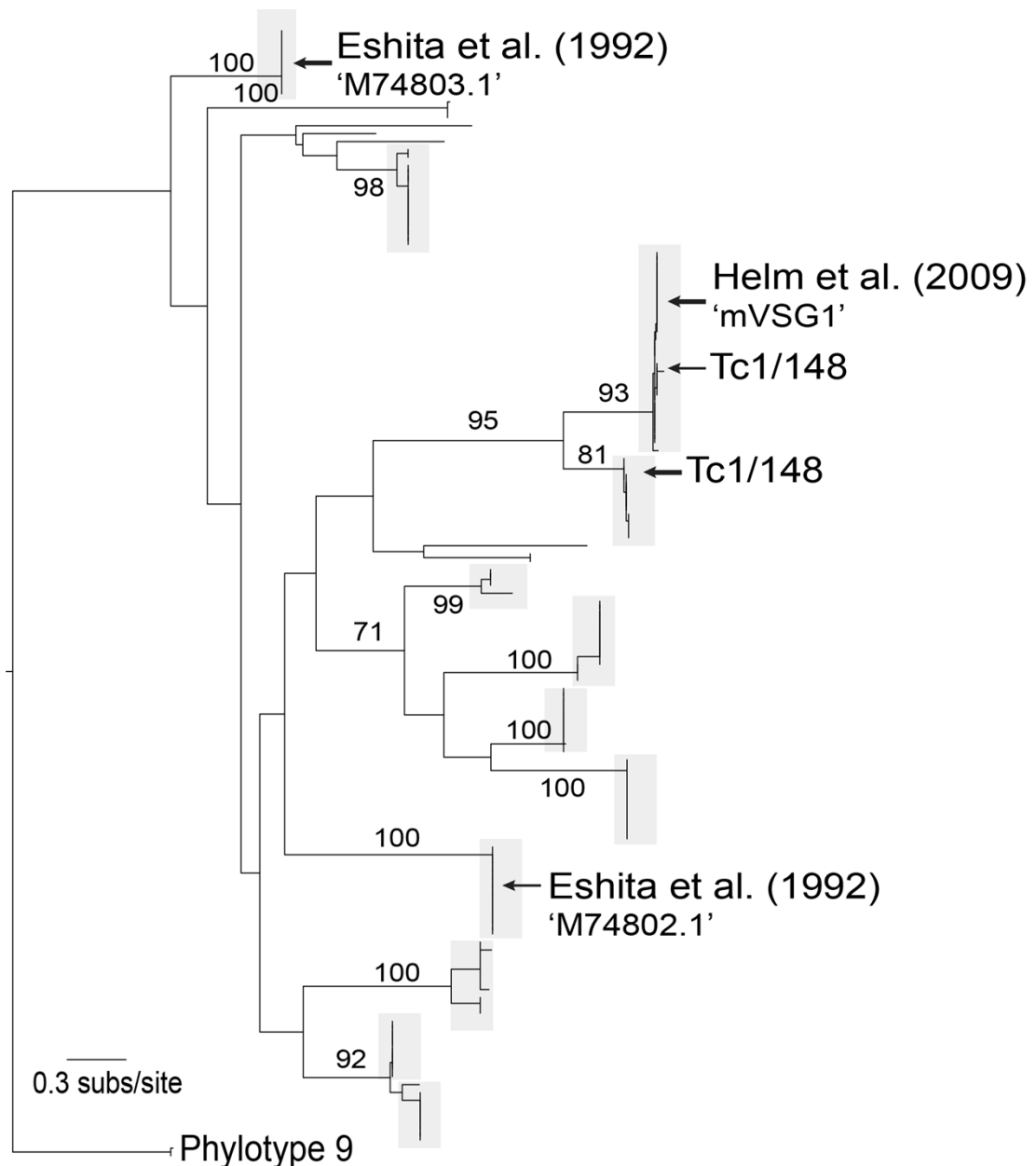


Figure 26 Maximum likelihood phylogeny of phylotype 8 estimated from protein sequences. Tree includes all the phylotype 8 sequences recovered from the isolates described in chapter 2; the two mVSG sequences described by Eshita et al. (1992); and mVSG1 described by Helm et al. (2009). Internal nodes are labelled with bootstrap values higher than 70 %. Arrows indicate positioning of the mVSGs described here and in the literature. Tree is rooted on phylotype 9.

Total protein was extracted from the 24 individual samples and sequenced by mass spectrometry. Resulting peptides were screened against a protein database of non-redundant VSGs, non-VSG Tc1/148 proteins and *Glossina sp.* proteins. After an initial analysis with PEAKS (Ma et al. 2003), 6 VSG peptides were observed in 2

samples (samples 1 and 10). These peptides corresponded to one protein group per sample. The VSG protein group of sample 1 contains 4 potential VSGs from phylotype 5, whilst sample 10 contains 2 VSG matches to phylotype 15. The proteomic evidence for sample 1 and 10 does not match either the transcriptomic profiles or the telomeric expression site profile. Both observations can be due to the low degree of confidence in the peptide search. Additionally, the lack of concordance between a proteomic or transcriptomic profile and the expression site profile may be insignificant for two reasons: first, as with *T. brucei*, the metacyclic expression site may be distinct from the bloodstream form expression site and therefore not included in the ES profile; second, if only one expression site is active at a time, then the profile of the inactive ES is irrelevant. In summary, the proteomic results are conflicting; therefore, the degree of post-transcriptional regulation and the composition of the mVSG coat of parasites from the same population remain unclear.

3.3.4 The expressed mVSG repertoires of metacyclic-enriched populations

Although epimastigotes do not express VSGs on the surface, they could transiently express them at the mRNA level. It is also not known how VSG transcription and expression occurs in pre-metacyclics, i.e. the stage between epimastigotes and fully differentiated metacyclics. To evaluate the effect of such transcripts in final VAPs, population enrichment for metacyclics was attempted. Parasites were released from the hypopharynx by tissue disruption and gentle shaking and stored on ice in PBS and glucose. As transcriptomes from individual flies showed good reproducibility and because enrichment would considerably decrease the amount of starting material for RNA and protein extraction, samples were pooled into two replicates of 60 and 71 flies each.

The composition of the population before and after enrichment was estimated by microscopy and cell cytometry, revealing that the non-enriched mouthpart parasite population was mostly composed of epimastigotes and other non-metacyclic intermediate forms (up to 82 %). After poly-L lysine slide adhesion enrichment, the percentage of metacyclics in the population increased by 30 %. After DE52-cellulose separation, there was a drastic decrease in total number of cells, but the population became predominantly composed of metacyclics (56 % and 76 %) (**Figure 27**). Yet,

the VAPs produced from the transcriptomes of both enriched methods show high replication (**Figure 28**) and are very similar to those obtained from non-enriched samples (**Figure 23, infection 3**), suggesting that the non-metacyclic forms present in the non-enriched samples from individual flies did not affect the number or composition of VSG transcripts recovered.

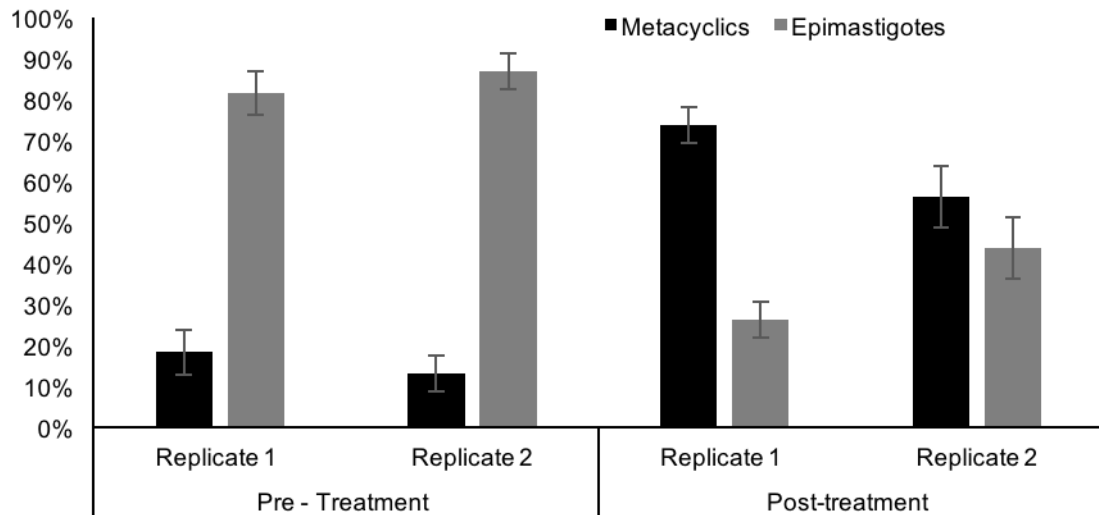


Figure 27 Trypanosoma populations before and after DE52-cellulose column separation (mean \pm σ). Bar charts show the enrichment of metacyclic populations after DE52-cellulose separation.

With the exception of Replicate 1 from the Poly-L lysine slide adhesion method, a high number of different VSG transcripts were recovered from the metacyclic enriched samples (N = 30, 165 152, and 235, respectively). However, the transcriptomes from individual samples recovered 373 distinct VSGs combined, which suggests that there were many VSGs common to the 24 individual samples and that the sequencing effort was sufficient to capture a true glimpse of the VSG transcriptome even when those are a small proportion of the total cells in the fly mouthparts.

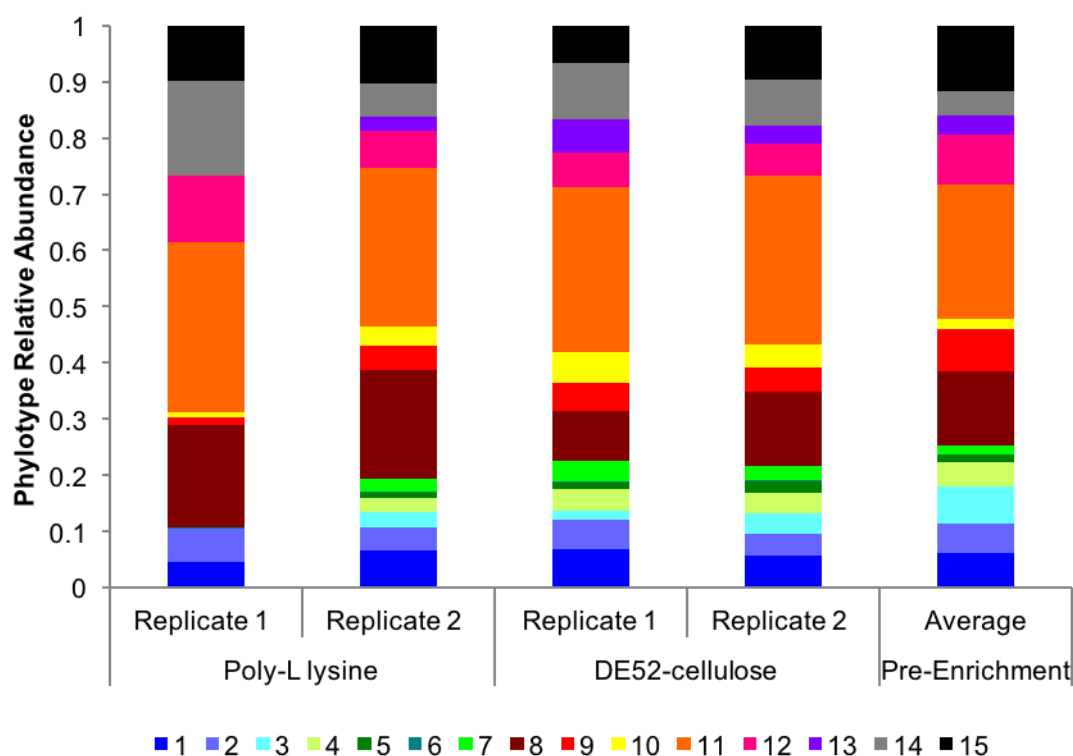


Figure 28 Transcriptomic Variant Antigen Profiles of trypanosomes extracted from pooled tsetse mouthparts after metacyclic-parasite enrichment. The number of VSG transcripts used in the VAPs are as follows: in Poly-L lysine slide adhesion method, 30 and 165 transcripts were recovered for replicate 1 and 2, respectively; in the samples from the DE52-cellulose column separation, there were 152 and 235 VSG transcripts in replicate 1 and 2, respectively.

The protein extracted from the metacyclic-enriched parasite suspensions obtained from DE52-cellulose separation was sequenced by mass spectrometry and searched for *T. congolense* VSGs, using the *Glossina sp.* proteome as host background. The discovery analysis revealed 5 protein groups, each group containing multiple VSGs from the same phylotype. The results suggest that the metacyclic population expresses phylotypes 1, 5, 3, 11, and 15, which remains conflicting with the transcriptomic results (**Table 8**). However, although the numbers of recovered peptides and their confidence value have increased in the enriched, pooled samples compared to the individual mouthparts, the sensitivity of the proteomic results remains low, prejudicing an accurate biological conclusion.

Table 8 Proteomic Results from mass spectrometry sequencing of metacyclic-enriched parasite suspensions obtained by DE52-cellulose separation. Data represents the number and confidence (-10lgP) of VSG peptides matches obtained for each replicate and for a ‘combined’ sample. ‘Combined’ relates to an independent search conducted on the merged data from the two replicates. Note that VSGs belonging to the same phylotype also share the same protein group in the proteomics search.

	VSG	Protein Group	Number of Peptides	-10lgP	Phylotype
Replicate 1	Tc14802288	1436	1	32.11	1
	Tc14808532	1436	1	32.11	1
	Tc14810144	1436	1	32.11	1
	Tc14814692	1436	1	32.11	1
	Tc14815124	664	2	38.78	15
Replicate 2	Tc14802288	1449	1	25.71	1
	Tc14808532	1449	1	25.71	1
	Tc14810144	1449	1	25.71	1
	Tc14814692	1449	1	25.71	1
	Tc14808333	517	3	53.91	11
	Tc14802502	1434	1	31.36	13
	Tc14805217	1434	1	31.36	13
	Tc14805420	1434	1	31.36	13
	Tc14810929	1434	1	31.36	13
	Tc14815124	426	2	47.22	15
Combined	Tc14802288	2064	1	24.67	1
	Tc14808532	2064	1	34.24	1
	Tc14810144	2064	1	24.67	1
	Tc14814692	2064	1	24.67	1
	Tc14892320	2064	1	24.67	1
	Tc14815193	1896	1	32.38	5
	Tc14802502	2053	1	30.13	13
	Tc14805217	2053	1	30.13	13
	Tc14805420	2053	1	30.13	13
	Tc14810929	2053	1	30.13	13
	Tc14815124	519	3	62.45	15

In summary, the results from this chapter show that the VAP can be applied to transcriptomic data to provide a fast and reliable metric of VSG expression. Additionally, the analysis of the mVSG repertoire of Tc1/148 suggests that the mVSG expression is reproducible between infections and populations, and may be epigenetically controlled through the developmental regulation of particular phylotypes.

3.3.5 Further attempts to detect mVSG by MS

As the proteomic results obtained lacked depth and confidence, a final fly infection was carried out with the exclusive purpose of protein extraction. Initially, a parasite sample from a pool of 106 mouthparts was thoroughly washed in PBS to reduce the signal from tsetse proteins. Although the MS analysis revealed only minor contamination from tsetse proteins, no VSG peptides were identified in the sample. Lastly, material from 37 infected mouthparts was washed as previously and used for GPI-PLC treatment. The latter aimed to enrich the sample for GPI-anchored proteins, such as the VSG, and soluble protein. The MS analysis revealed only one VSG peptide, representing a single VSG from phylotype 2.

3.4 Discussion

The main aim of this chapter was to adapt the VAP for the analysis of transcriptomic data to facilitate studies of VSG expression. During that process, I investigated the composition and variability of the expressed mVSG repertoire of metacyclic *T. congolense* 'savannah' Tc1/148 in the tsetse fly.

Transcriptomic VAPs provide the ability to characterise and quantify differences in VSG expression during experiments. By analysing RNAseq data, VAPs can be adjusted for transcript abundance, providing a description of VSG expression. To achieve this, the *de-novo* assembly process used in the genomic profiling was replaced by reference mapping and transcript abundance estimation. Although transcripts can be assembled *de-novo* using software like Trinity (Haas et al. 2013), transcript mapping to a reference genome is a more sensitive method, particularly when there are good reference genomes available. Even though read mapping to a different strain may be suboptimal in variable parts of the genome, such as the subtelomeres, the transcript abundance estimation software can perform *de-novo* transcript discovery using the mapped RNAseq data. Yet, if *de-novo* transcript assembly is preferred by the user, it can be easily incorporated as a preliminary step, before read mapping back to the transcriptome and transcript abundance estimation. After transcript abundance estimation, variant antigen profiling is performed as described in Chapter 2, using Hidden Markov Model search. Finally, VAPs are adjusted for transcript abundance to achieve accurate representation of the VSG expression. This step is important because it allows the discrimination amongst phylotypes according to the weight each transcript represents in the total expressed VSG repertoire, rather than their frequency. This is particularly useful to differentiate between phylotypes composed of multiple low-abundance transcripts and phylotypes with few high abundance transcripts.

In this chapter, the concept of the transcriptomic VAP was applied to investigate the composition and variability of the mVSG repertoire expressed by metacyclic parasites in the tsetse fly. Previous analyses of the *T. congolense* mVSG repertoire were either performed in early mammal infections under the assumption that mVSG are still expressed up to one week after transmission (Crowe et al. 1983) or *in vitro* (Eshita et al. 1992). Although cultured metacyclic parasites might express similar mVSGs to those in the tsetse fly (Luckins et al. 1981), there is evidence in *Plasmodium falciparum* that vector passage can change gene expression, and that

var gene expression is reset after one mosquito passage (Bachmann et al. 2016). Therefore, this work was performed from tsetse mouthpart infections.

The first challenge of this approach was to produce transcriptomes from the mouthparts of individual flies. Analyses of salivary gland transcriptomes (or sialome) have been reported in the literature for over 10 years. Sialomes obtained using Expressed Sequence Tag (EST) libraries have been produced for a range of insects, including *Glossina morsitans morsitans* (Alves-Silva et al. 2010), but also the dengue and yellow fever vectors *Aedes aegypti* (Ribeiro et al. 2007) and *Aedes albopictus* (Arcà et al. 2007), and *Triatoma infestans*, which transmits Chagas disease (Assumpção et al. 2008). Furthermore, transcriptomes using the microarray technique have been produced for the malaria vector *Anopheles gambiae* to assess differential expression before and after blood feeding (Das et al. 2010). The more advanced RNAseq technique has been used recently for the analysis of sialomes of other insects, such as the western flower thrips (*Frankliniella occidentalis*) (Stafford-Banks et al. 2014).

All these studies focused on the non-infected insect salivary glands and yet all were performed with pooled samples ranging from 15 (for the large *T. infestans* EST library) to 300 (for the small *F. occidentalis* RNAseq library). RNAseq analysis of the parasites colonising the salivary glands of an insect has only been done very recently for *T. brucei* (Savage et al. 2016). This work presented a global transcriptome of the different parasite stages during development in the tsetse fly and was based on the isolation of pools of midgut, proventriculus, and salivary glands. As *T. congolense* bypasses the salivary glands, migrating from the proventriculus to the proboscis (mouthparts), this chapter presented mouthparts transcriptomes rather than sialomes. Yet, as their size is comparable, the challenges of isolating enough RNA from the salivary glands of a single fly are similar to those of the proboscis. This chapter shows that it is possible to produce transcriptomes from a single mouthpart.

In the study of Savage et al. (2016), metacyclic parasites were isolated from blood. Flies with mature salivary gland infections were fed, and the blood remaining on the mouthparts was collected, pooled and the parasites isolated by DE52 anion exchange. This approach, whilst necessary for the producing an accurate metacyclic transcriptome, precludes the use of single flies. As the work presented in this chapter had the main purpose of mVSG analysis, a perfect separation of

metacyclic from the remaining parasite forms present in the mouthparts was not essential. The inclusion of the whole proboscis with all its parasitic contents ensures enough RNA can be isolated for deep sequencing. The insect reads can be easily removed by read mapping to the tsetse genome and the remaining parasite forms do not affect the VSG analysis because only metacyclic parasites contribute to the mVSG expression profile (as shown by the metacyclic-enriched VAPs).

Trypanosome gene expression is post-transcriptionally regulated with the sole exception of mVSGs in *T. brucei*, which are transcribed from a transcriptionally regulated monocistronic expression site (Lenardo et al. 1986; Graham & Barry 1995). *T. brucei* metacyclics are thought to express a single mVSG, presumably covering their surface (Tetley et al. 1987; Ramey-Butler et al. 2015). From my data, I cannot infer whether *T. congolense* single parasites express a single mVSG, or multiple. For such a conclusion, *T. congolense* metacyclics would either have to be engineered to express fluorescently-labelled mVSGs (if the mVSG expression cassette was known) or through single-cell proteomics. However, it does show that the metacyclic parasite populations within individual flies express very similar VSG profiles. This in itself is a marked difference to *T. brucei*, where the expressed mVSG repertoire is subject to change during tsetse transmission, as shown by Barry et al. (1983) who detected changes in expression of 3 metacyclic VATs during sequential tsetse transmission of a laboratory parasite clone.

Another difference in *T. congolense* metacyclic transcriptomes is that the mVSGs are rarely within the 200 most abundant transcripts. The low mVSG transcript abundances could refer to various factors, such as inefficiency in sequencing due to vector RNA contamination, high percentage of epimastigotes in the sample, and the fact that metacyclic cells show a drastic decrease in gene transcription. Yet, when Savage et al. (2016) sequenced *T. brucei* populations from tsetse salivary glands, they observed that mVSG were amongst the 20 most abundant transcripts, even though there was a high percentage of epimastigotes represented, for example, by the high abundance of brucei alanine rich protein (BARP). In contrast, the results of this chapter show a wide range of distinct mVSG being transcribed in the population (average of 79.2 ± 31.8 per sample), at a level approximately ten-fold lower than in *T. brucei* (Savage et al. 2016). Therefore, it is possible that, unlike *T. brucei*, *T. congolense* mVSGs are not monoexpressed or that the surface of the *T. congolense* metacyclic parasites are not fully coated with VSGs.

Whilst mVSG expression in *T. congolense* is different from *T. brucei*, perhaps they agree on their synchronicity. Ramey-Butler et al. (2015) showed that in *T. brucei* multiple mVSGs are expressed simultaneously in the population and that expression onset is rapid and synchronous rather than gradual. Likewise, this chapter suggests that *T. congolense* metacyclic populations may also express multiple VSGs. Furthermore, assuming that differences in flies and parasite stocks affect infection kinetics, as the VSG expression profiles remain similar between infections, it is plausible that expression is also synchronous.

The proteomic results obtained in this work failed to convince and disagreed with the transcriptomes. In the enrichment data, the proteomic results are stronger and seem to agree in the expression of phylotype 1, 5, 11, 13, and 15, but still disagree with the transcriptomes, particularly in the case of phylotype 5 (which is barely expressed in the transcriptome), and the absence of phylotype 8, which is highly expressed in the transcriptomes. Although it is possible that in *T. congolense* part of the VSG expression control is post-transcriptional, this has never been reported and strongly disagrees with the current knowledge for *T. brucei*, where mVSG expression is exclusively transcriptionally regulated (Lenardo et al. 1986; Graham & Barry 1995; Barry et al. 1998). Another possibility would be that there is a lag between mRNA transcription and protein expression, which, in the event of a highly dynamic metacyclic surface, would account for the disparities between transcriptomes and proteomes. However, the observation of Ramey-Butler et al. (2015) that the delay between mVSG RNA transcription and protein expression in *T. brucei* is negligible, and the fact that the multiple transcriptomes presented in this chapter come from different infections and potentially from parasites at distinct developmental stages and yet retain a strong reproducible VAP, suggests otherwise.

At the transcriptomic level, the results presented here suggest that metacyclic parasites express a distinct, limited, and reproducible set of mVSGs. The extent of this reproducibility across strains remains unclear, but it is maintained in Tc1/148 having survived a full transmission cycle. As phylotype 8 remains consistently over-represented in the metacyclic transcriptomes, and their members are always among the most abundant VSG transcripts, preferential expression of this phylotype could be postulated. In fact, this is corroborated by an earlier study of mVSG expression in *T. congolense* ILNaR2 (Eshita et al. 1992), which identified two mVSGs, through cDNA expression library in phages and Northern blotting, both belonging to

phylotype 8 (M74802.1 and M74803.1) (**Figure 26**). More recently, the most abundant mVSG observed through sequencing of an expression sequence tag library from *T. congolense* IL3000 metacyclics cultured *in vitro* also belonged to phylotype 8 (mVSG1) (Helm et al. 2009) (**Figure 26**). The described phylotype 8 mVSGs have orthologues in various strains and cluster across the whole extent of the tree, not forming their own independent group, suggesting that if there is functional differentiation and/or developmental regulation, this is characteristic of the phylotype rather than individual genes. Therefore, there is evidence from distinct strains pointing towards phylotype 8 being preferentially expressed in metacyclic parasites, and so possibly developmentally regulated.

As discussed in chapter 2 (page 69), developmental regulation and functional differentiation amongst variant antigens is present in African trypanosomes as well as other protozoan parasites. For example, functional differentiation amongst *Plasmodium falciparum* var genes is well represented by the *var2csa* gene. This gene, unique for retaining orthology across *P. falciparum* strains, has apparently evolved an invariant function in regulating the expression of other family members (Ukaegbu et al. 2015). Hence, it is reasonable to hypothesize that subtle yet important functional differences underlie the maintenance of VSG phylotypes in *T. congolense* and cause the specific metacyclic profile observed in this study.

It remains to be seen whether phylotype 8 is restricted to metacyclics and whether it is enriched in natural fly infections, however, all of these observations contrast with what is known in other antigenically variable vector-borne organisms. In *T. brucei*, the mVSG repertoire, although limited, progressively changes over time both in natural infections and sequential laboratory tsetse transmissions of the same parasite clone (Barry et al. 1983). In *P. falciparum*, the var gene expression radically changes following a single mosquito passage (Bachmann et al. 2016), indicating that the specific var genes subset being expressed by a parental line is completely converted during the development in the mosquito, resulting in a novel subset being expressed when entering a naïve host. In *P. chabaudi*, vector passaging not only alters *cir* (chabaudi interspersed repeats) expression in the erythrocytic cycle, but also leads to virulence attenuation, related to the broad activation of most subtelomeric variants (Spence et al. 2013). If the pattern of *T. congolense* mVSG expression is reproducible in nature, then the preferential expression of phylotype 8 (and perhaps others) indicates a form of developmental regulation that could potentially be exploited in vaccine design. Metacyclic VSGs have long been targets

for vaccine development because they are the first point of contact with the host immune system. Although their instability in *T. brucei* has impaired any progress towards a vaccine, if the same phylotypes are over-represented in all *T. congolense* strains, vaccination may be plausible in *T. congolense*.

Another key question that arises from the results is how certain phylotypes are repressed and favoured during the metacyclic stage. In this study, the analysis of the physical characteristics of the various phylotypes did not reveal striking differences that could explain differences in transcription. However, the growing evidence for extensive epigenetic regulation of variant gene expression in parasitic protozoa suggests that epigenetic regulators, such as chromatin modifiers, writers and chaperones, are likely candidates for the regulation of VSG expression (Duraisingh & Horn 2016). Evidence for the role of heterochromatin in controlling allelic exclusion in *T. brucei* is abundant (Figueiredo & Cross 2010) and involves a multitude of players, such as Histone 1 and 3 (Pena et al. 2014; Povelones et al. 2012; Alsford et al. 2012), H3.V, DOT1B (Figueiredo et al. 2008; Batram et al. 2014; Reynolds et al. 2016; Schulz et al. 2016), the chromatin remodellers ISWI (Stanne et al. 2015) and nucleoplasmin-like protein (NLP) (Narayanan et al. 2011), the histone chaperone FACT (Denninger et al. 2010), and the nuclear lamin NUP1 (DuBois et al. 2012). Additionally, alternative epigenetic machineries, such as SUMOylation, appear to be involved in VSG activation (López-Farfán et al. 2014).

In this light, further experiments targeting over and under-represented phylotypes in the metacyclic stage with the aim to identify and characterise the potential epigenetic signature of metacyclic VSG expression in *T. congolense* can help explain the mechanisms behind any programmed VSG expression.

3.4.1 Conclusions

This chapter shows that the VAP can measure differences in VSG expression and it exposed an unexpected degree of reproducibility in antigenic expression. The fact that mVSG expression is stable, reproducible and partly distinct from bloodstream VSG expression is a key difference from the mechanisms described for *T. brucei*, urging further studies of VSG expression dynamics in *T. congolense*. The systematic over-representation of particular phylotypes during the metacyclic stage, coupled with evidence for rare recombination between phylotypes, can revive the idea of using metacyclic VSG to vaccinate against *T. congolense*. Foremost, progress towards this objective will require further experiments to characterise the stability of the phenotype observed in this study over longer periods, amongst isolates of different origins, and in the context of mixed infections, which are a common feature of natural settings.

Chapter 4. Characterisation of conserved telomeric structures in *T. congolense*

4.1 Introduction

In *T. brucei*, VSG expression occurs at telomeric expression sites. Initial attempts to resolve the structure and regulation of the *T. brucei* BES revealed that the ES were polycistronic units where the VSG and a minimum of 7 non-VSG genes were transcribed under regulation from the same promoter (Kooter et al. 1987; Johnson et al. 1987; Shea & Van der Ploeg 1988). Through cloning experiments, these units were shown to vary between 45 and 60 kb in length and α -amanitin resistance assays surprisingly revealed that transcription was driven by RNA polymerase I (Shea et al. 1987; Alexandre et al. 1988; Shea & Van der Ploeg 1988; Pays et al. 1989). The non-VSG genes co-transcribed in the ES were named ESAGs. However, homologous genes were also found to be transcribed in procyclic stage, where VSG transcription is absent, and not by RNA polymerase I (Gibbs & Cross 1988; Pays et al. 1989; Graham & Barry 1991). These results were further supported by the observation that ESAG4 was homologous to adenylate cyclases, suggesting that ESAGs might be derived from multi-copy gene families conserved in the wider genome (Pays et al. 1989).

In more recent years, research has focused on revealing the fine detail of the ES. Subtelomeres and telomeres are difficult to sequence through conventional short-read sequencing methods, due to their unpredictable length and highly repetitive composition. Thus, other approaches have been especially developed to allow the study of the VSG system and the resolution of the architecture of all telomeric ES, such as bacterial-associated chromosome (BAC) cloning (Berriman et al. 2002) and transformation-associated recombination (TAR) in yeast (Larionov et al. 1996). The latter yields better outcomes, achieving successful isolation of complete chromosome ends and the generation of stable clones of 17 BES from *T. brucei* Lister 427 (Becker et al. 2004). For this method, the TAR cloning vector pEB2 was modified to contain a stable yeast telomere and the *T. brucei* 427 dominant expression site promoter as a recombinational target. This strategy resulted in >85 % successful clones as linear yeast-associated chromosomes (YACs) with no

unexpected recombination. Further work on the library of TAR clones allowed the detailed study of 14 BES groups and the consequent revealing of the ESAG set (Hertz-Fowler et al. 2008), adding to the 4 BES from 4 different strains successfully sequenced before (Berriman et al. 2002). This analysis confirmed that *T. brucei* BES are 10-50 kb long polycistronic regions with a canonical structure comprising an upstream promoter, a collection of ESAGs, a 70 bp repetitive sequence, and the terminal VSG, with individual differences lying in the ESAG sets (**Figure 29**). Nonetheless, all were located within 60 kb from the end of the chromosome and the expressed VSG within 2 kb of the telomeric repeat.

T. brucei has 12 ESAGs: ESAG 9, 10, and 12 are facultative, whereas the remaining are present in all Lister 427 BES sequenced to date (**Figure 29**). ESAG1 has been described as a *T. brucei*-specific gene and is located adjacent to the subtelomeric end of the 70 bp-repeat. ESAG2 is a gene family adapted from an ancestral lineage of b-type VSG, whose closer relatives are Fam16 *T. congolense* VSGs (Jackson et al. 2012). ESAG3 and ESAG5 are membrane-associated proteins with divergent homologues in the core chromosomes of *T. congolense* and *T. brucei* (Hertz-Fowler et al. 2008). ESAG4 is an adenylate cyclase, although not orthologous to any adenylate cyclases of *T. congolense* (Pays et al. 1989; Jackson et al. 2013). ESAG6 and ESAG7 are found almost exclusively at the BES as tandem pairs (Jackson et al. 2013). They encode transferrin receptors (Salmon et al. 1994), which have largely expanded in *T. congolense*, but are mostly located in the subtelomeres and not associated with the VSG (Jackson et al. 2012). ESAG8 is a putative nuclear DNA-binding protein which binds a protein involved in mRNA stability (Hoek et al. 2000; Hoek & Cross 2001; Hoek et al. 2002); ESAG9 is a stumpy stage-specific molecule of unknown function, with no homologues in trypanosomatids other than weak sequence similarity to the MASP proteins of *T. cruzi* (Barnwell et al. 2010), a family of mucin glycoproteins also of unknown function (Bartholomeu et al. 2009). ESAG10 is a folate transporter identical to core homologues. ESAG11 is a modified invariant surface glycoprotein (Jackson et al. 1993), and ESAG12 is a *T. brucei*-specific gene of unknown function (Hertz-Fowler et al. 2008). With the exception of ESAG8, all ESAGs have cell-surface roles and, with the exception of ESAG10, all are thought to be derived from multi-copy gene families distributed throughout the core genome. Furthermore, only ESAG1 and ESAG12 seem to have evolved *de-novo* in the *T. brucei* genome, whilst all the remaining have distant homologues in *T. congolense*, *T. vivax*, or *T. cruzi* (Jackson et al. 2013).

All *T. brucei* ESAGs derive from multi-copy gene families present in the core chromosomes or subtelomeres. However, following independent recruitment to the ES, they have adapted to their new chromosomal location and evolved concertedly due to the many pressures for sequence exchange and reorganisation in the telomeres (Jackson et al. 2013). Thus, in this chapter I take the view that ESAGs are genes that, whilst derived from conserved, multi-copy gene families have independently diversified to the expression site, resulting in distinct isoforms present exclusively in the expression site context.

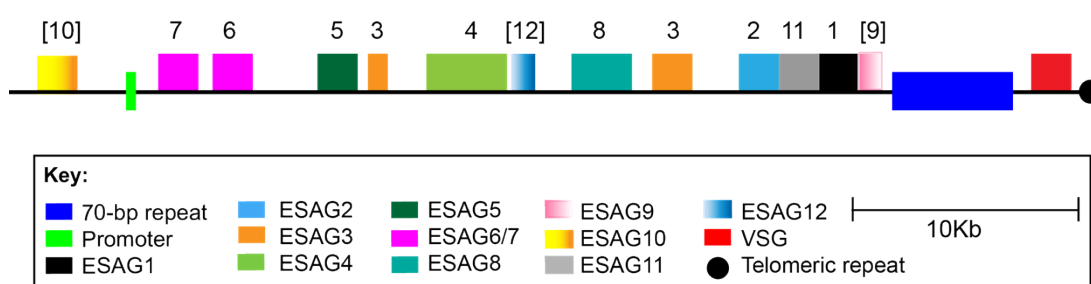


Figure 29 The consensus structure of the *T. brucei* Lister 427 bloodstream expression site (BES). Features are scaled and colour coded according to key. Gradient-shaded ESAGs in square brackets represent facultative ESAGs. Adapted from Hertz-Fowler et al. (2008).

Research in *T. congolense* is much less developed. *T. congolense* VSGs are thought to also be expressed from telomeres, but the structure of its ES is yet to be described. With the emergence of long-read genome sequencing technologies, such as the Single Molecule, Real-Time (SMRT) cell sequencing from Pacific Biosciences (Rhoads & Au 2015), recovering and describing telomere-associated regions is easier. Telomeres can be sequenced in reads as long as 60 kb, which can often represent the telomere-associated structures of the genome in single reads without the need for assembly.

The IL3000 actively-transcribed VSG and its surrounding sequence was cloned, but never published, by John Donelson in 2005. The sequences of four clones are publicly available in GenBank under the accession numbers HE578911, HE578912, HE578913, and HE578914. They show the active VSG immediately upstream the telomeric repeat, preceded by two non-coding regions, the transposable element ingi, a range of non-VSG genes (i.e. RNase A, cathepsin-B, zinc-finger protein, hypothetical protein), and the 369 bp repeat previously described as a feature of the

mini-chromosomes (Majiwa et al. 1986; Gibson et al. 1988) (**Figure 30**). Furthermore, two telomeric contigs from megabase chromosomes were recovered from the *T. congolense* genome sequencing (Jackson et al. 2012), which share similar features. Whilst these sequences provided preliminary data for the analysis conducted in this chapter, they represent a poor sample of not only the megabase chromosomes, but of the telomeric context of *T. congolense*. In this study, two strains of *T. congolense* ‘savannah’ were sequenced with the SMRT cell technology to reveal telomere-associated structures. This chapter aims to:

1. Describe *T. congolense* telomere-associated structures and compare them to *T. brucei* to investigate the presence of a canonical expression site in *T. congolense*.
2. Detect and characterise telomere-associated non-VSG genes to evaluate whether they are orthologous or analogous to *T. brucei* ESAGs.
3. Profile the VSGs found in the telomeres to investigate whether specific phylotypes are preferentially recruited to the telomeres.
4. Explore the role of recombination in telomere evolution to understand whether they are under similar pressures as *T. brucei* ES.
5. Understand the role of mini-chromosomes in VSG diversity and expression.

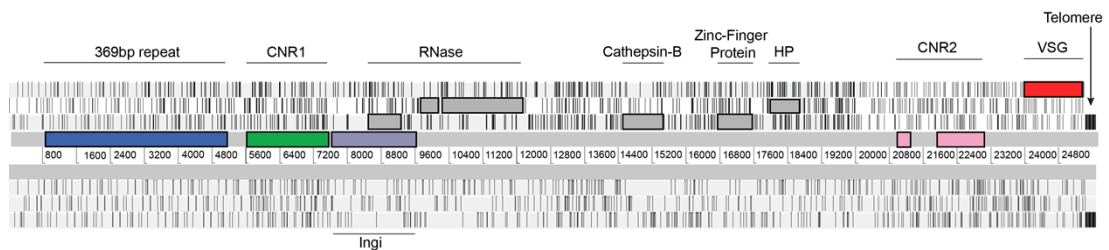


Figure 30 The genomic context of the *T. congolense* IL3000 actively-transcribed VSG, or the ‘Active VSG ES’. The actively-transcribed VSG of IL3000 is preceded by a stretch of 369 bp repeats, CNR1, the transposable element ingi, the RNase gene, a cathepsin B gene, a zinc-finger protein gene, a gene encoding a hypothetical protein, and CNR2. The VSG gene is the most telomere-proximal element.

4.2 Methods

4.2.1 Parasite stocks and culture

T. congolense 'savannah' Tc1/148 (MBOI/NG/60/1-148) (1/148) (Young & Godfrey 1983) mouse blood stabilates were obtained from the Department of Parasitology of the Liverpool School of Tropical Medicine, UK. A frozen stabilate of 1ml of infected mouse blood (20 % glycerol and parasitaemia of 10^3 parasites/ml) was thawed and mixed with ten volumes of defibrinated horse blood. The sample was used to infect 100 male tsetse flies by pouring 4mls of blood onto three distinct spots on a feeding tray covered with a feeding membrane. Flies were allowed to feed for 11 min at 27 °C and in the dark. Flies were killed by decapitation at day 10 post-infection and midguts dissected in sterile conditions. Infected midguts were incubated with modified Eagle's medium (MEM)-based modified differentiating trypanosome medium (DTM) (10 % foetal bovine serum, 2 mM L-glutamine, 10 mM L-proline) and 0.5 mg/mL penicillin/streptomycin to prevent growth of midgut bacteria on 96-well plates. Parasite suspensions were serial-diluted 5 times at a 1:2 ratio of parasite: medium to remove fly tissue and incubated at 27 °C, 5 % CO₂. Cultures were passaged at a 1:10 ratio first into 12-well plates and then into 25 ml flasks with fresh medium when parasitaemia reached 10^6 cells/ml. Cultures continued to grow until a total of 1.2×10^9 cells were obtained.

4.2.2 DNA extraction

High molecular weight DNA was extracted from 1.2×10^9 cells by phenol: chloroform protocol. Cells were centrifuged at 1,500 g for 10 min and washed in 10ml cold PBS. Cells were centrifuged at 1,500 g for 10 min and supernatant was discarded. Pellet was resuspended in 500µl PBS and incubated with 6ml TELT buffer (1.5 M LiCl anhyd, 50 mM Tris-HCl pH8.0, 62.5 mM EDTA pH8.0, 4 % Triton-X) at room temperature for 5 min. In a fume hood, 7 ml of 1:1 phenol: chloroform was added and mixed by inversion for 5 min or until an emulsion was formed. The solution was centrifuged at >3,000 g for 5 min and the aqueous solution collected in a 50 ml tube. The remaining phases were discarded. Two volumes of ethanol were added to the aqueous solution, mixed by inversion, incubated on ice for 10 min and centrifuged at 4,000 g for 20 min. The supernatant was discarded by gentle decantation and the pellet was washed in 2 volumes of freeze-cold 70 % ethanol. The solution was

centrifuged at 3,000 g for 5 min and the supernatant decanted. The pellet was left to air dry at 70 °C for 5 min and then re-dissolved in 600 µl TE50 (10mM Tris-HCl pH8.0, 50mM EDTA pH8.0). 150 µm/ml of RNase A was added to the resuspended pellet and incubated for 1 hour at 37 °C. Subsequently, 300 µg/µl of Proteinase K was added to the solution and incubated for 2 hours at 50 °C. After the incubation period, 600 µl 1:1 phenol: chloroform was added in a fume hood and mixed by inversion for 5 min. The solution was centrifuged at > 3,000 g for 5 min and aqueous fraction collected in a 1.5 ml tube. To the aqueous solution, 1 volume of isopropanol and 0.1 volumes of 3 M sodium acetate (NaOAc) were added. The solution was centrifuged at 1,500 g for 15 min at 4 °C. The supernatant was discarded and the pellet was washed in 1 ml ice-cold 70 % ethanol. The pellet was left to air dry until no ethanol was visible. Finally, the pellet was left to re-dissolve in TE50 (2 µl/10⁷ cells) at 4 °C overnight, without pipetting or mechanical disturbance.

Tc1/148 DNA output (105 µg) was quantified with Qubit fluorometric dsDNA quantitation (dsDNA HS Assay Kit) (Life Technologies, UK). DNA quality was checked on a Nanodrop (Thermo Scientific, UK) (A₂₆₀/A₂₈₀ = 1.82 and A₂₆₀/A₂₃₀ = 1.78) and integrity checked on a 0.5 % agarose gel run at 30 V for 16 hours at 4 °C, using the 1 kb DNA extension ladder (Invitrogen, UK).

4.2.3 Long-read genomic DNA library preparation and sequencing

DNA was purified with 1x cleaned Ampure beads (Agencourt) and the quantity and quality was re-assessed using Qubit fluorometric dsDNA quantitation (dsDNA HS Assay Kit) (Life Technologies, UK) and Nanodrop (Thermo Scientific, UK). Fragment analyser (Bioanalyzer, Agilent, UK) with a large fragment high sensitivity genomic kit (Agilent Genomics, UK) was used to determine the average size of the DNA. 10 µg of sample was used without further shearing.

DNA was treated with Exonuclease V11 at 37 °C for 15 min. The ends of the DNA were repaired as follows: samples were incubated for 20 min at 37 °C with damage repair mix supplied in the SMRTbell library kit (Pacific BioSciences, USA). This was followed by a 5-min incubation at 25°C with end repair mix. DNA was cleaned using a 1:1 volume ratio of AMPure beads (Agencourt, UK) and 70 % ethanol washes. DNA was ligated to adapter overnight at 25°C. Ligation was terminated by incubation at 65 °C for 10 min followed by exonuclease treatment for 1 hour at 37

°C. The SMRTbell library was purified with a 1:1 volume ratio of AMPure beads (Agencourt, UK). The quantity of library and therefore the recovery was determined by Qubit assay and the average fragment size determined by Fragment analyser as before. Size selection was performed on Sage blue pippin prep using 0.75 % agarose cassette/S1 marker and the size collected was between 15,000 bp and 50,000 bp. The final SMRTbell library was recovered and the DNA damage repaired as previously done at the start. SMRTbell libraries were annealed to sequencing primer at values predetermined by the Binding Calculator (Pacific Biosciences, USA) and a complex made with the DNA Polymerase (P6/C4 chemistry). The complexes were bound to Magbeads and this was used to set up the required number of SMRT cells for each sample. For Tc1/148, 20 kb Libraries were sequenced on the PacBio SMRT sequencer, using 360-min movie times and 7 cells.

4.2.4 Genome assembly

Single pass reads generated on the PacBio® SMRT sequencer (Pacific BioSciences, USA) were assembled using the Hierarchical Genome Assembly Process 3 (HGAP3) (Chin et al. 2013), under default conditions and a predicted genome size of 34 Mb. This software uses an Overlap-Layout-Consensus algorithm to generate a polished de novo assembly of large genomes. Other assembling softwares are available for this purpose, but HGAP3 is recommended by the sequencer manufacturer PacBio and is implemented within the SMRT Portal. The polished assembly contained 536 contigs (n50 = 421,740 bp), assembled from 201,878 reads of 14,594 bp on average. The genome has been deposited in ENA/GenBank under the accession number NHOR01000000.

The IL3000 genome was produced at the Centre for Genomic Research (University of Liverpool) on the PacBio SMRT sequencer and given to us as a draft genome containing 1,415 contigs (n50 = 156,211 bp) for the sole purpose of VSG ES analysis.

4.2.5 Genome annotation

For both strains, assembled contigs were annotated using the web server Companion (Steinbiss et al. 2016), using RATT (Otto et al. 2011) on species mode

to transfer relevant annotation from *T. brucei* 927, with *ab initio* gene finding using AUGUSTUS (Stanke et al. 2004) with a score threshold of 0.7 to make gene prediction more sensitive.

4.2.6 Contig selection

Identification of telomere-containing contigs

The first step in the investigation of the telomeric context of *T. congolense* was to identify and retrieve genomic contigs containing a string of 3 or more telomeric repeats at their ends. This was performed using Repeat Masker (<http://repeatmasker.org>) on the polished assembly, which searches DNA sequences for repeats and low complexity DNA sequences and provides detailed annotation on their location and composition. Thus, all the contigs with identifiable strings of telomeric repeats were selected and manually curated. Contigs with telomeric repeats not located at the ends were discarded and considered sequencing or assembly errors.

Identification of Contigs Containing IL3000 telomeric features

To check that the features identified in the IL3000 telomeric contigs from John Donelson and the sequencing project (Jackson et al. 2012) were specific to telomere-associated contigs, the polished assemblies were screened by sequence similarity search [BLASTn and tBLASTx (Altschul et al. 1990)]. BLAST was also used to search for conserved regulatory regions and ESAGs. The conserved non-coding regions (CNR) described in section 4.3 ('Results'), were identified from the preliminary data (369 bp repeat, CNR1 and CNR2) and from repeated observations of conserved open reading frames (CNR3 and CNR4) in the BLAST output. A significance threshold (E-value) of 10^{-4} was applied and all regions manually checked by ACT (Carver et al. 2005) and multiple sequence alignment.

4.2.7 Contig annotation

The features identified by the sequence similarity searches were annotated into the respective contigs using Artemis (Rutherford et al. 2000).

4.2.8 Multiple Sequence Alignment

Nucleotide sequences of conserved non-coding regions of *T. congolense* Tc1/148 were aligned with ClustalW (Larkin et al. 2007) and manually curated. The 126 CNR1 sequences produced an alignment of 3632 nucleotides. The conserved Non-Coding Region 2 (CNR2) produced an alignment of 37 sequences and 176 nucleotides. The conserved Non-Coding Region 3 (CNR3) produced an alignment of 15 sequences and 199 nucleotides. The conserved Non-Coding Region 4 (CNR4) produced an alignment of 21 sequences and 148 nucleotides. Amino acid sequences of conserved coding regions found in multiple telomere-containing contigs (Fam15, Fam53, DEAH-box RNA helicase, cathepsin B) were aligned with ClustalW (Larkin et al. 2007). The Fam15 alignment was comprised of 217 sequences from *T. brucei* and *T. congolense* IL3000 and Tc1/148 of 435 amino acids. The Fam53 protein alignment consists of 173 sequences from *T. brucei*, *T. congolense* IL3000 and Tc1/148, *T. vivax* and *T. cruzi* and has 294 amino acids. The DEAH-box RNA helicase alignment contains 50 sequences from *T. brucei* and *T. congolense* IL3000 and Tc1/148 of 1089 amino acids. The cathepsin B alignment has 23 sequences from *T. brucei* and *T. congolense* IL3000 and Tc1/148 of 347 amino acids.

4.2.9 Phylogenetic estimation

To investigate orthology between *T. brucei* and *T. congolense* ESAGs and their relationship with family members outside the ES context, phylogenies were estimated for each of the ESAGs using protein sequences. Phylogenies were estimated from protein sequence alignments with maximum likelihood following automatic model selection (Lefort et al. 2017) using PHYML v3.0 (Guindon & Gascuel 2003) and RaXML (Stamatakis 2014). Robustness was assessed with 100 bootstrap replicates.

4.2.10 Comparison of tree topology

When the *T. brucei* telomeric ES were described by Hertz-Fowler et al. (2008), it was found that the tree topologies of different loci were distinct across the ES. This showed that the phylogenetic signal along the ES was variable, and thus indicated

that sequence recombination was a driver of ES evolution. To evaluate whether the same applied to *T. congolense*, the differences in phylogenetic signal between different loci of the ES were calculated by evaluating the significance of the differences in likelihood values between the optimal tree and the constrained tree. ML phylogenetic trees were estimated for each CNR1 and CNR2-4 using the Tamura-Nei model (Tamura & Nei 1993) in MEGA7 (Kumar et al. 2016), resulting in 4 phylogenies. Under the same methods, phylogenetic topologies were constrained to the topology of the optimal CNR1 tree, resulting in 3 phylogenies. Likelihood values recorded for the phylogenies of CNR2-4 and the differences between constrained and unconstrained topologies were calculated. The significance of such differences was evaluated using the Shimodaira-Hasegawa test (Shimodaira & Hasegawa 1999) in RaxML (Stamatakis 2014). Likelihood comparisons require comparable trees to have equal taxon sets, thus only contigs containing both of the features being compared were included (i.e. CNR1 and CNR2; CNR1 and CNR3; CNR1 and CNR4; CNR3 and CNR4). Because the number of contigs containing CNR2 and CNR3 or CNR2 and CNR4 was below the minimum threshold for phylogenetic estimation ($N = 4$), the phylogenetic signal between them was not measured.

4.2.11 Recombination and selection tests

To understand whether the DEAH-box RNA helicase family expansion was a telomeric-unique phenomenon (and therefore linked to their transposition to that genomic context), the role of recombination in the expansion was investigated. This was performed by predicting breakpoints with the Genetic Algorithm for Recombination Detection (GARD) (Kosakovsky Pond et al. 2006). GARD was run using the REV model, under the AICc information criterion. The Kishinoi-Hasegawa (KH) test was applied to test for rate heterogeneity to account for the effect of significant topological incongruences. The role of selection in sequence evolution was assessed with three site-level selection tests. These were the Fixed Effects Likelihood (FEL) to directly estimate dN/dS ratios (Pond & Frost 2005); the Random Effects Likelihood (REL) to infer selection pressures using an empirical Bayes approach and model $\omega(dN/dS)$ ratios at individual sites based on a pre-defined distribution; and the Fast Unbiased Bayesian Approximation (FUBAR) to estimate ω based on Bayesian Inference using a MCMC routine (Murrell et al. 2013). One branch-level test (BRL) based on an empirical Bayes approach was also applied to

infer selection pressures in individual phylogenetic lineages (Kosakovsky Pond et al. 2011).

4.3 Results

To investigate whether *T. congolense* has conserved telomere-associated structures, as observed in *T. brucei*, long-read SMRT sequencing of two strains, Tc1/148 and IL3000, was performed to allow detection of contigs containing telomeric repeats at the ends. This approach also allowed the description of non-VSG genes found in the telomeres and the investigation of the role of recombination in telomeric sequence evolution. As the context identified in both genome sequences were consistent with the sequences from the clones of the IL3000 of the actively-transcribed VSG, I will treat the telomere-containing contigs and their structure as *T. congolense* VSG ES.

4.3.1 The telomere-containing contigs

In my Tc1/148 genome sequence, 153 contigs contained a string of telomeric repeats at their end. These included 25 complete mini-chromosomes, containing one telomeric repeat at each end and a long complex 369 bp repeat in the middle. A complex repeat homologous to the 369 bp repeat previously described in *T. congolense* mini-chromosomes (Moser et al. 1989) was found in 80 % of the contigs and 66 % had a non-coding region that resembles a transcription promoter.

The 369 bp complex repeat was considered the proximal boundary of the ES because it occupies the centre of complete mini-chromosomes, being flanked by CNR1 at each end. This suggests the presence of two identical ES, one at each telomere. Furthermore, when the 369 bp complex repeat is found in the megabase chromosomes, it is the most proximal conserved region identified, separating the ES from the subtelomere.

Three other conserved non-coding regions were identified: CNR2 was found in 31 % of the contigs and also in the 'IL3000 ES'. CNR3 and CNR4 were found in 8 and 12 % of the contigs, respectively. Full-length VSGs were found in 50 % of contigs, occasionally adjacent to other VSGs and VSG pseudogenes. Coding sequences found in the telomeres included 2 contigs with a cathepsin B (1 %), 7 contigs (5 %) with transferrin receptors (Fam15), 3 contigs (3 %) with ESAG3-like genes (Fam53), 17 (11 %) with a DEAH-box RNA helicase, 16 contigs (10 %) with RHS genes, and

a range of low-copy number hypothetical proteins. The transposable element Ingi was found inside the ES in 15 contigs (2 %).

In the IL3000 genome, the scenario was very similar. Six complete mini-chromosomes were identified, with a structure identical to those in Tc1/148, and two contigs were assigned to megabase chromosomes 9 and 11. Of the 128 contigs with a string of telomeric repeats at their end, 127 contained the 369 bp repeat (99 %) and 122 the CNR1 (95 %). The CNR2-4 were found in 77 (60 %), 25 (20 %), and 30 (23 %) contigs, respectively, and VSGs in 51 contigs (40 %). Cathepsin B was found in one contig (1 %), Fam15 was found in 5 contigs (4 %), and 7 contigs (5 %) contained DEAH-box RNA helicases.

4.3.2 *T. congolense* has canonical telomeric structures

Taking into account both strains, a consensus canonical telomeric structure was inferred and shown in **Figure 31**. This consists of the 369 bp repeat, CNR1, the VSG and CNR2-4. These are the features that are more abundant and exist always in the same relative position.

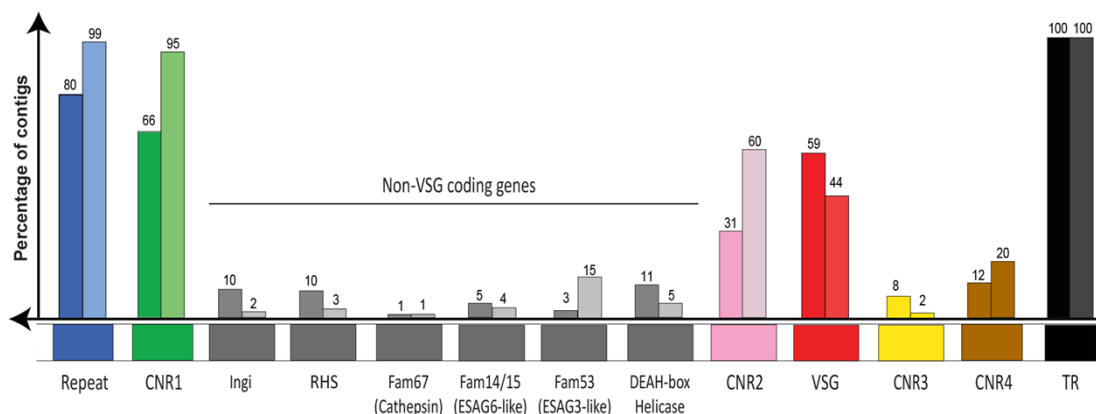


Figure 31 The consensus structure of the telomere-associated structures in *T. congolense*. Conserved structures are colour-coded and named and their frequency in the telomere-containing contigs given by the bar chart.

The 369 bp repeat

The 369 bp repeat was found in 80 and 99 % of contigs, respectively, always close to the telomere and exclusively in the context of the ES. This repeat is highly repetitive and has an AT-rich complex sequence (66.71 %±0.37). **Figure 32** shows

a multiple sequence alignment of the 369 bp repeat from the Tc1/148 and IL3000 genome sequences. For clarity, only 50 sequences from each strain were used. The figure highlights the high nucleotide conservation, mostly greater than 95 %, across contigs and between strains. In Tc1/148 and some sporadic contigs of IL3000, the repeat is 359 bp rather than 369 bp, due to an insertion between nucleotides 241 and 250 in IL3000 (marked in **Figure 32** by dashed box). Occasionally, and in IL3000 only, the repeat is palindromic.

The 369 bp repeat has been previously described as a feature of mini-chromosomes, analogous to the 177 bp repeats in *T. brucei* (Moser et al. 1989; Sloof et al. 1983). This is consistent with our results in both strains. A southern blot gel run by Simon D'Archivio at the University of Nottingham confirms that the repeat is present in the mini-chromosomes (**Figure 33**), but it does not show evidence for the 369 bp repeat presence in megabase chromosomes. This contradicts the finding of the 369 bp repeat in the megabase chromosome 5 of Tc1/148 (1.57 bp in length, spanning the subtelomere and the core), and its presence in the IL3000 original genome sequencing project (which was a back-end library of the megabase chromosomes) (Jackson et al. 2012). Both these studies associate the 369 bp repeat with megabase chromosomes and suggest it is not exclusive to the small chromosomes, but rather to the telomeric context.

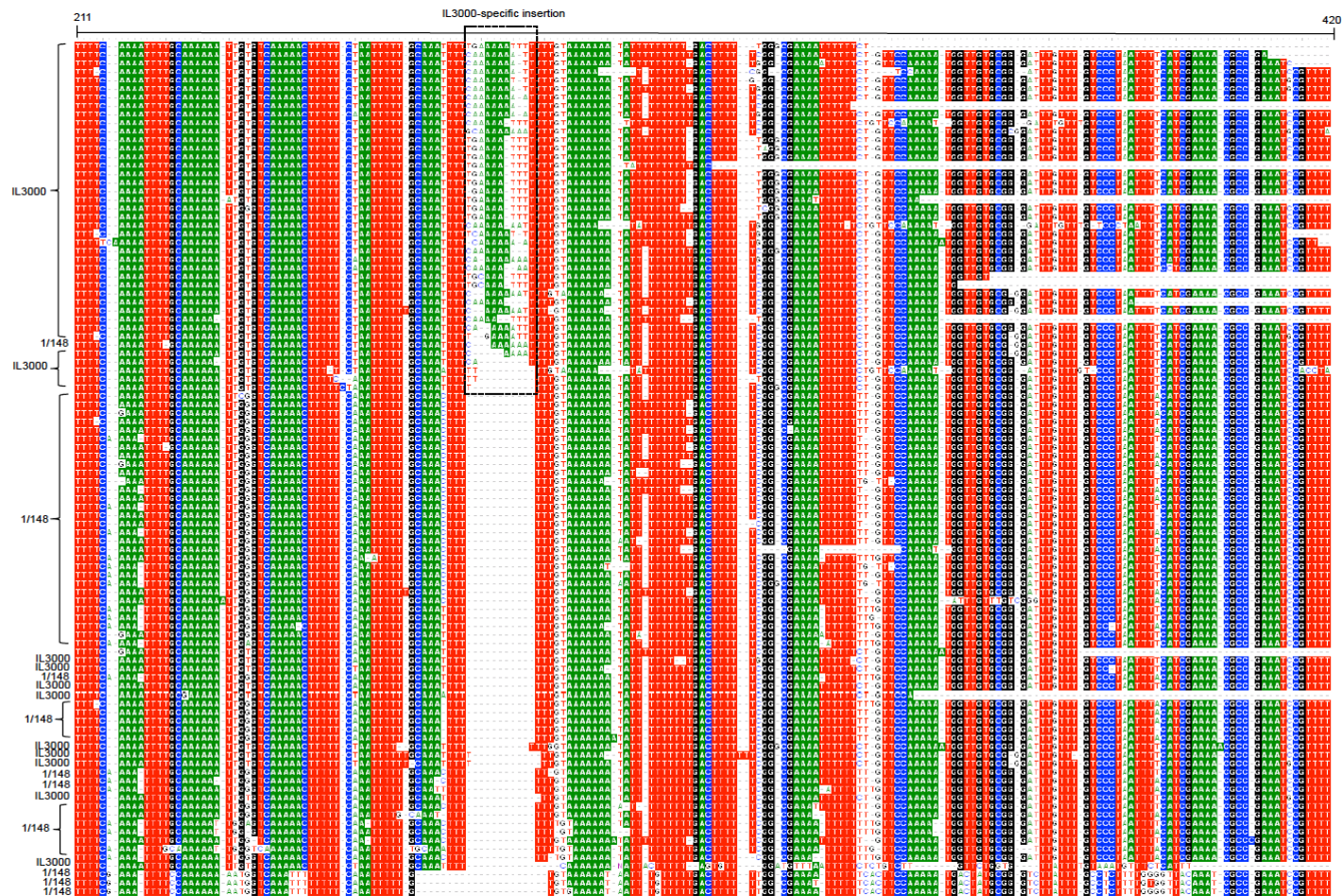


Figure 32 Multiple sequence alignment of the 369 bp repeat in IL3000 and Tc1/148. Strain identity is marked on the left. IL3000-specific insertion is denoted by dashed box. Shade represents nucleotide conservation $\geq 95\%$.

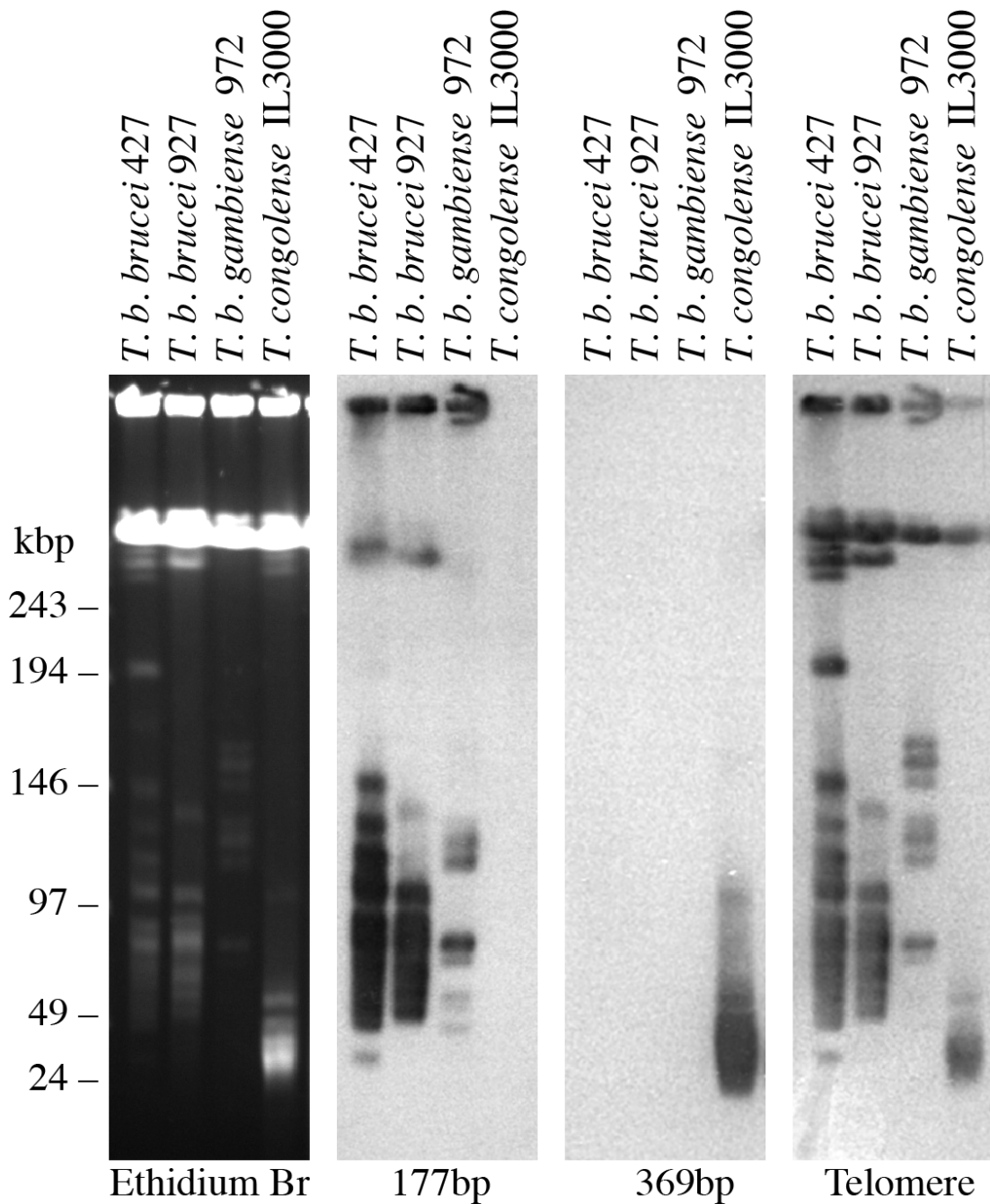


Figure 33 The karyotype of *T. brucei* spp. and *T. congolense*. Southern Blot gel shows staining with ethidium bromide, hybridization to labeled mini-chromosome satellite repeats (177 bp in *T. brucei* spp. and 369 bp in *T. congolense*, and hybridization to labeled telomeres (D'Archivio, S., unpublished).

CNR1

CNR1 is a conserved non-coding region of a variable size up to 3584 bp, but with a consistent conserved region common to all contigs (65 % sequence identity) and an ultra-conserved core of 302 nucleotides between residue 1893 and 2195 sharing 87

% identity. **Figure 34** shows the nucleotide variation (entropy) and the frequency of each residue across the alignment of all CNR1 sequences. The graph shows nucleotide conservation is highest (i.e. lowest entropy) around residue 2129; the most variable regions locate just before residue 1445 and around residue 1813. In terms of residue frequency across all sequences, the core part of the alignment (i.e. between nucleotides 1445 and 2813) is present in most sequences, whereas the beginning and end of the alignment are less common, suggesting a high length variation in the region. This structure is only found within the ES context.

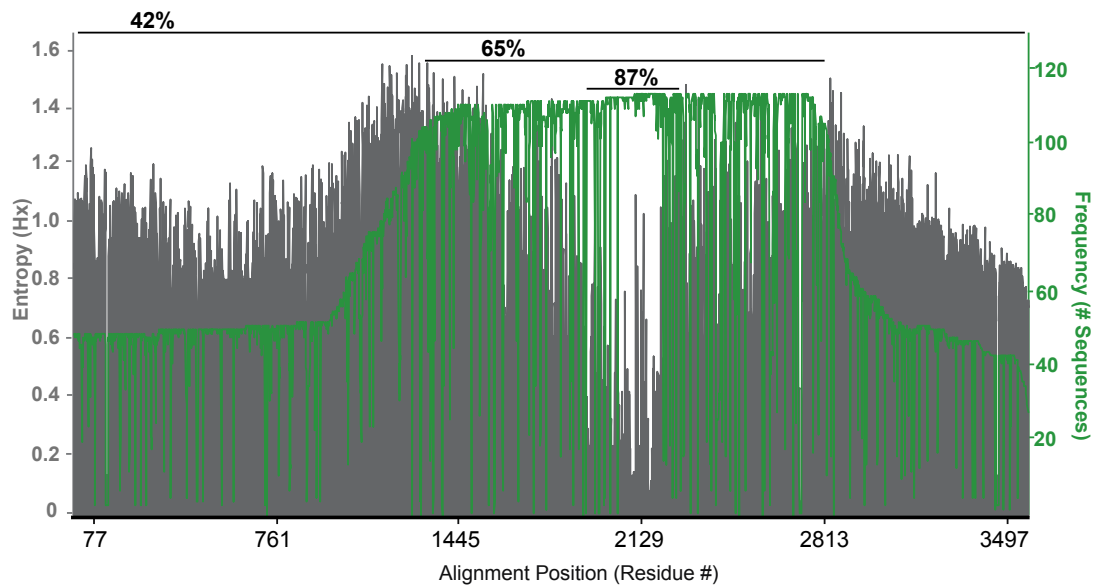


Figure 34 CNR1 sequence variation and frequency in Tc1/148. Graph shows the Shannon entropy of the multiple sequence alignment of the CNR1 region in grey (left axis) and the frequency of each residue across the multiple contigs in green (right axis). Top percentages represent sequence identity at each region.

VSG

In Tc1/148, VSGs were found in a telomeric context in 90 contigs. In the vast majority, they represented the most telomere-proximal coding sequence. To determine whether VSGs were randomly assigned to the telomeres or whether particular phylotypes were consistently absent, all VSG from the telomeres were retrieved and analysed. Results show that they derive from both Fam13 and Fam16 and that most, but not all, phylotypes are represented (**Table 9**). Phylotypes 6 and 9 were not found at the telomeres in either Tc1/148 or IL3000. The genomic and ES VSG profiles of each strain correlate very poorly (IL3000 $R^2 = 0.07$; Tc1/148 $R^2 =$

0.002). Additionally, whilst the genomic VSG profiles of IL3000 and Tc1/148 somewhat correlate ($R^2 = 0.59$), their ES profiles do not ($R^2 = 0.11$). Nonetheless, the difference between profiles was not significant either between individual phylotypes (student t-test, $p > 0.05$ for all phylotypes) or between complete VAPs (Poisson regression model, $p > 0.05$). Therefore, it is still possible that the differences in repertoire observed between the genome and the ES are due to chance alone.

Table 9 Variant antigen profiles of the genomic VSG and the telomeric VSG (ES) from IL3000 and Tc1/148. Values are shaded according to key. Blue represents low frequency and red represents high frequency of a given phylotype.

Phylotype	Genome IL3000	ES IL3000	Genome 1/148	ES 1/148
1	0.073	0.094	0.072	0.345
2	0.040	0.094	0.032	0.034
3	0.049	0.188	0.032	0.034
4	0.018	0.313	0.004	0.152
5	0.060	0.000	0.007	0.014
6	0.055	0.000	0.005	0.000
7	0.054	0.000	0.202	0.021
8	0.025	0.188	0.004	0.055
9	0.012	0.000	0.006	0.000
10	0.060	0.000	0.014	0.014
11	0.142	0.031	0.130	0.069
12	0.100	0.000	0.133	0.083
13	0.061	0.000	0.054	0.110
14	0.078	0.000	0.039	0.021
15	0.172	0.094	0.265	0.048

CNR2

CNR2 is conserved non-coding region composed of 180 bp repeats found exclusively in a telomere-associated context. It is always located between the CNR1 and the VSG at conserved distances. In Tc1/148, CNR2 is associated with a telomere in 36 contigs. In those, in 75 % of the cases (27), it is located an average of 4,229 (± 770) bp upstream of the telomere. In the remaining 25 % of the cases (N

= 9) it is located it is further away, i.e. 20,500 (\pm 3,347) bp upstream of the telomere. The region is also associated with a VSG in 20 contigs: in 16 of those it locates 2,498 (\pm 944) bp upstream of it and in 4 of them is positioned 16,023 (\pm 2,784) bp upstream. Whilst the overall nucleotide similarity can be as low as 40 %, CNR2 is compositionally consistent and various blocks of highly conserved nucleotides (> 70 %) are visible (**Figure 35**), such as a 5'-GTTGATT-3' string between residues 78 and 85 and a 5'-GCTCTCT-3' string between positions 102 and 108 of the alignment. Furthermore, its relative positioning to the telomere and the VSG is conserved in both strains.

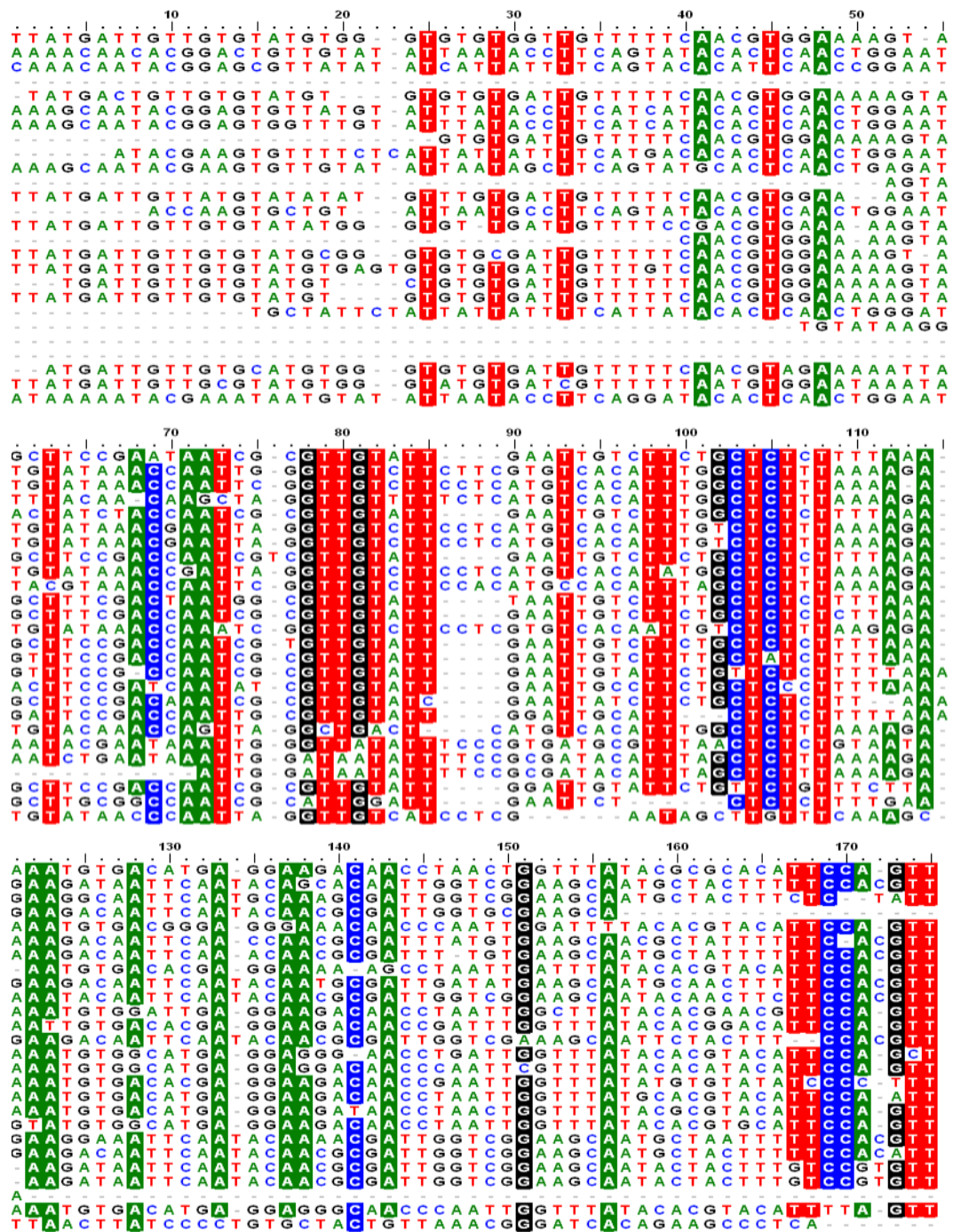


Figure 35 Multiple sequence nucleotide alignment of CNR2. Top sequence is from IL3000 active ES; remaining sequences were retrieved from the Tc1/148 ES. Shade represents nucleotide conservation $\geq 70\%$.

CNR3

CNR3 is a highly conserved 200-nucleotide sequence present exclusively in the context of the VSG ES (). CNR3 can be found in 12 contigs in Tc1/148 always downstream of the VSG. On average, CNR3 locates 1,047 (\pm 280) bp downstream of the VSG and 536 (\pm 174) bp upstream of the telomere. CNR3 was identified in IL3000 as the pseudogene TcIL3000_04880, but it is most likely a conserved non-coding region. There is no evidence for gene expression in any of the available datasets for *T. congolense* [EST library data (Helm et al. 2009), the bloodstream form IL3000 transcriptome (Wellcome Trust Sanger Institute pre-release (<https://www.ebi.ac.uk/arrayexpress/experiments/E-ERAD-440/>), and all the epimastigote/metacyclic transcriptomes produced for chapter 3]. It is a *T. congolense*-specific sequence with no other gene orthologue anywhere in the reference genome, the GC content is higher than the average for *T. congolense* coding sequences, and there are multiple internal stop codons in the sequence.

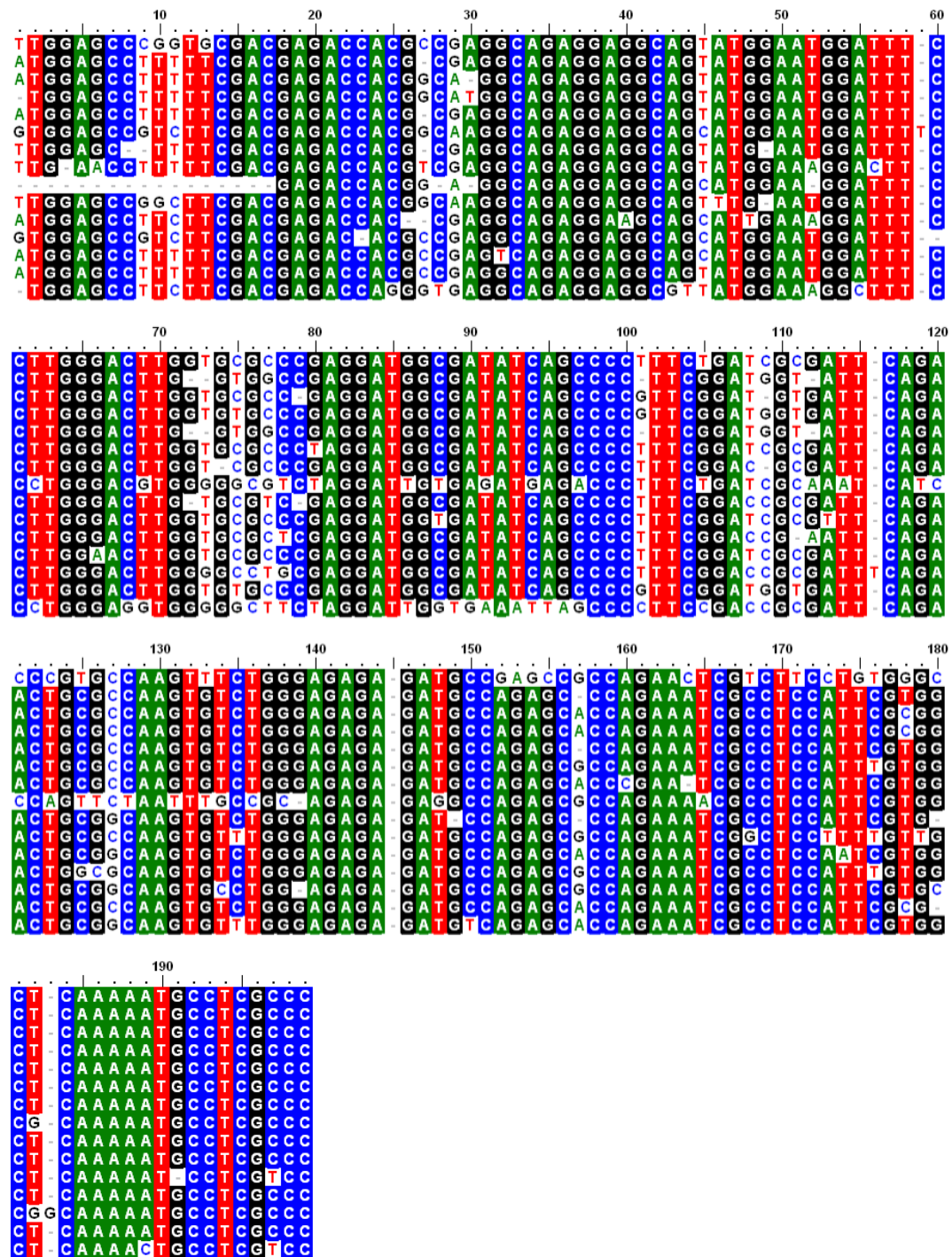


Figure 36 Multiple sequence nucleotide alignment of CNR3. Top sequence is from IL3000 genome (TcIL3000_0_04880); remaining sequences were retrieved from the Tc1/148 ES. Shade represents nucleotide conservation $\geq 70\%$.

CNR4

CNR4 is a 148 bp region highly conserved towards the 3' end (**Figure 37**). In Tc1/148, it was found in 18 contigs, two of which have two CNR1 sequences and thus were not used for positioning analysis. Of the 16 contigs with a single CNR1, CNR4 was always found associated with both a VSG and a telomere. In all instances, it locates 1,386 (\pm 483) bp downstream of the VSG and in 15/16 instances 458 (\pm 207) bp upstream of the telomere. In the only exception, it locates 12,234 bp upstream of the telomere, but still within the same distance from the VSG as the others. CNR4 was first identified in the IL3000 genome as the pseudogene TcIL3000_0_12610, but when putting it in the context of the remaining copies, it becomes more likely to be a non-coding region. Compositionally, CNR4 does not resemble a coding sequence, its translation in any frame results in heavily pseudogenic sequences, there is no orthologue in the IL3000 genome, there is no evidence for its expression in any of the available datasets [EST library data (Helm et al. 2009), the bloodstream form IL3000 transcriptome (Wellcome Trust Sanger Institute pre-release (<https://www.ebi.ac.uk/arrayexpress/experiments/E-ERAD-440/>), and all the epimastigote/metacyclic transcriptomes produced for chapter 3], and the GC content is higher than average for *T. congolense* coding sequences (60 % vs. 51 %).

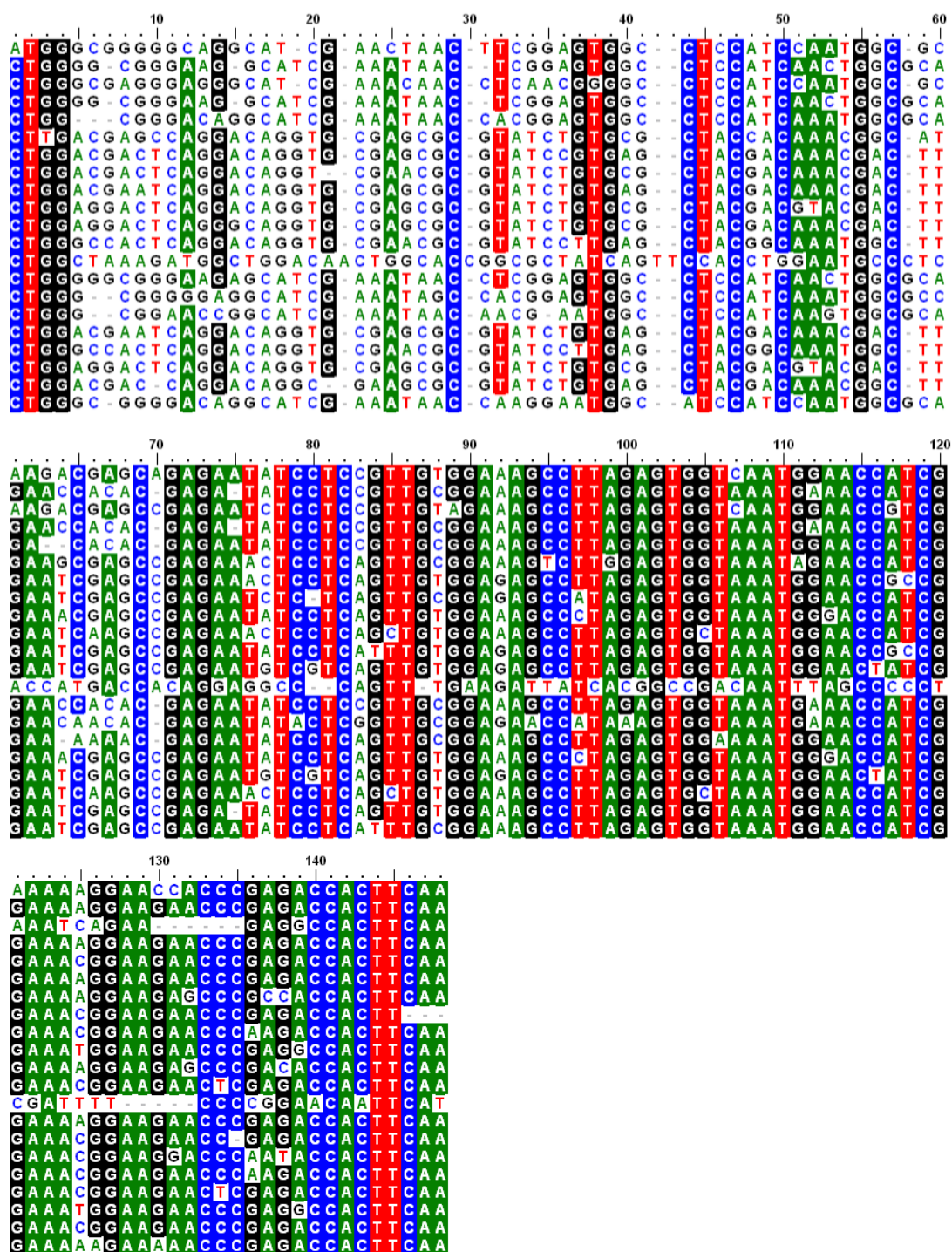


Figure 37 Multiple sequence nucleotide alignment of CNR4. Top sequence is from IL3000 genome (TcIL3000_0_12610); remaining sequences were retrieved from the Tc1/148 ES. Shade represents nucleotide conservation $\geq 70\%$.

4.3.3 *T. congolense* telomere-associated genes lack evidence for sequence adaptation to the telomere

Although not all ESAGs are essential for ES activation, the *T. brucei* ESAG repertoire, as a whole, is a necessary and constitutive feature of the BES (Becker et al. 2004; Hertz-Fowler et al. 2008). To understand whether the same principle applied to *T. congolense*, the coding sequences found within the telomeric context were analysed. The first striking difference between species is that the non-VSG coding sequences are not part of the canonical structure, but rather scarce and variable in position. There are coding regions that resemble those present in the *T. brucei* ES, such as the transposable element ingi, Fam15, and Fam53, yet they are not orthologous to those *T. brucei* ESAGs.

Fam15

Fam15 is a transferrin-receptor gene family present in *T. congolense* and *T. brucei*, but not in *T. vivax*. In 2013, Jackson et al showed that this family evolved from VSGs after the separation of *T. vivax* from the remaining African trypanosomes. Fam15 in *T. brucei* is comprised of ESAG6 (GPI+), ESAG7 (GPI-) and a single GPI+/GPI- tandem pair (Tb927.7.3250/3260) at a strand-switch region on chromosome 7. Therefore, ESAG6/7 are almost exclusively located at the BES and always in tandem pairs. In *T. congolense* IL3000, Fam15 has expanded (45 genes compared to 23 in *T. brucei*) and most copies of both GPI-positive and GPI-negative versions of the gene are found throughout the subtelomeres. In the telomeric context, Fam15 was found in 7 contigs in Tc1/148 and in 5 contigs in IL3000 (**Figure 38**). These genes are homologous to the copies found in the subtelomeres, but they are not necessarily arranged in tandem pairs. The phylogeny of Fam15 shows that although homologous, ESAG6/7 and *T. congolense* TFR genes are paraphyletic and therefore not orthologous to each other. This corroborates the findings of Jackson et al. (2013) that ESAG6/7 are a *T. brucei* specific innovation for the BES.

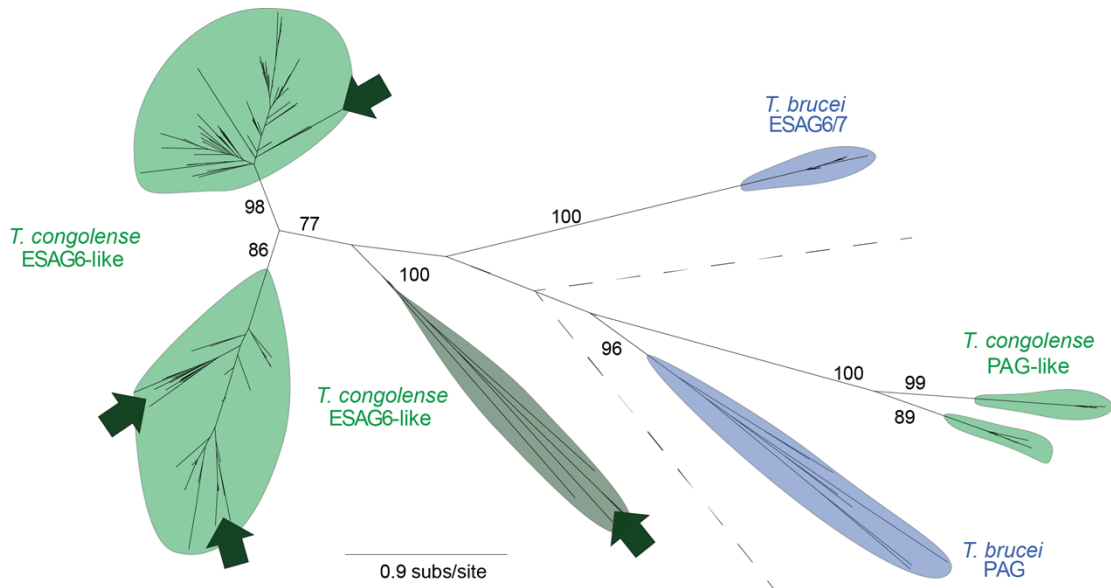


Figure 38 Consensus maximum likelihood phylogeny of the transferrin receptors protein sequences from *Trypanosoma congolense* and *Trypanosoma brucei*. The phylogeny was estimated with PHYML with a JTT+ Γ model and 100 bootstrap replicates. Terminal nodes are coloured by species according to key; bootstraps higher than 70 % are shown in the internal nodes. Green arrows show the position of *T. congolense* telomeric genes. Dashed line separates Fam14 (Procyclic-associated genes (PAG) and PAG-like genes) from Fam15 (ESAG6/7 and ESAG6-like genes).

Fam53

Fam53 encodes ESAG3 and ESAG3-like proteins in all trypanosomatids. However, in *T. brucei*, this family has expanded considerably in the subtelomeres; some members have transposed to the BES to become involved in VSG expression. ESAG3 is the only gene with flexible positioning within the ES and is often associated with recombination breakpoints, which might indicate a role of this gene family in BES recombination and/or VSG switching (Hertz-Fowler et al. 2008). Despite this expansion, *T. brucei* retains further subtelomeric copies, divergent in sequence to telomeric ESAG3-like genes ('GRESAG3'). These genes have co-orthologues in *T. congolense*, *T. vivax* and *T. cruzi* (**Figure 39**).

In *T. congolense*, only 5 copies of ESAG3-like genes were found at the telomeres, all homologous to the genes located at the subtelomeres. In IL3000, 19 telomere-associated contigs contained Fam53 members; all of these genes clustered with the Tc1/148 ESAG3-like genes from both the subtelomere and the telomere. However, in the subtelomere, two distinct genes were found to cluster separately from this

clade. The amino acid sequence of these genes is more similar to those of GRESAG3. Together, these findings suggest that, like *T. brucei*, *T. congolense* has also expanded ESAG3-like sequences. Despite this analogous situation, *T. congolense* and *T. brucei* are not orthologous, forming two distinct clades in the phylogeny (**Figure 39**).

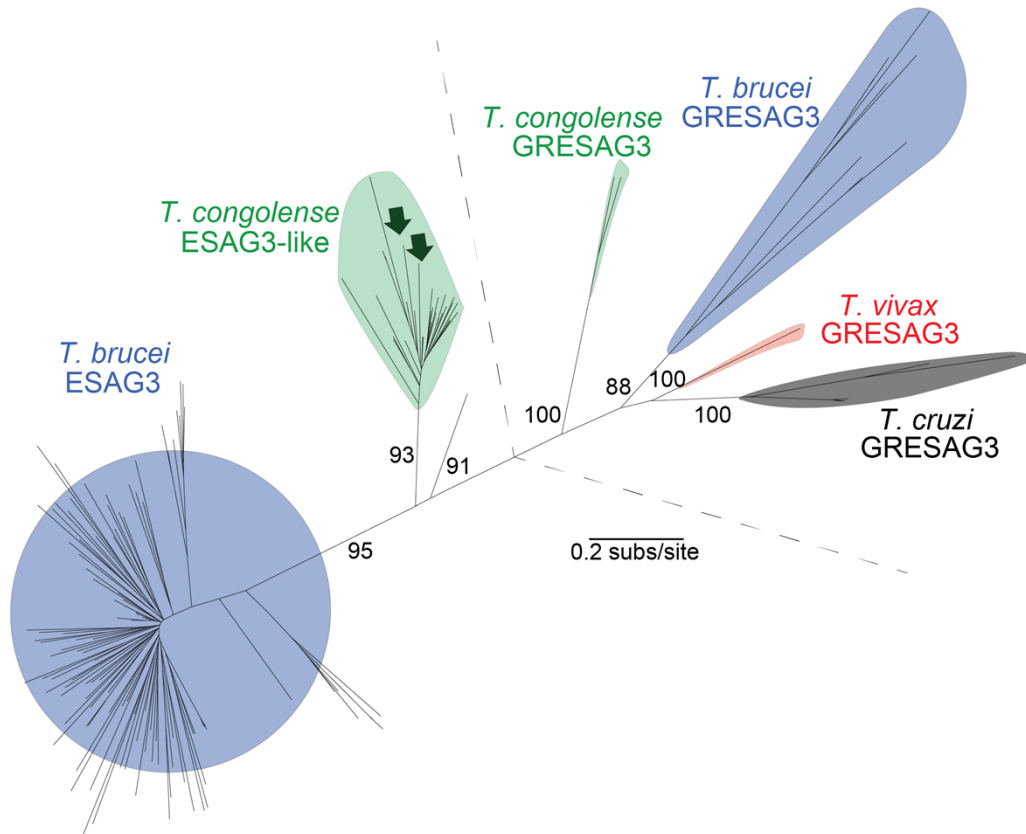


Figure 39 Consensus maximum likelihood phylogeny of Fam53 protein sequences from African Trypanosomes. The phylogeny was estimated with PHYML with a WAG+ Γ model and 100 bootstrap replicates. Clades are colour-coloured by species; internal nodes are labelled with bootstrap percentages higher than 70 %. Green arrows show the position of *T. congolense* telomeric genes. Dashed line separates ESAG3 and ESAG3-like genes from GRESAGs.

Whilst the data suggest that there are no orthologues to *T. brucei* ESAGs in *T. congolense*, there may be *T. congolense* specific genes that fulfil the criteria of an expression site associated gene. Such ESAGs could have originated from the same process, but recruited from different gene families. The other non-VSG coding

regions found distributed in the ES were cathepsin B, DEAH-box RNA helicases, and a range of hypothetical proteins.

Cathepsin-B

Cathepsin-B is a family of cysteine proteases that is single-copy in *T. brucei* and *T. vivax*, but has expanded in *T. congolense* (Mendoza-Palomares et al. 2008). Cathepsin-B is essential for *T. congolense* survival, being implicated in lysosomal protein degradation and immunogenicity (Mendoza-Palomares et al. 2008). One copy of cathepsin-B was found in the IL3000 actively transcribed VSG ES. This copy was also found in the new IL3000 assembly in a similar context. In Tc1/148, two telomeric contigs harbour cathepsin-B, all containing the canonical structures described above and the flanking genes present in the active ES of IL3000 (i.e. RNase A, zinc-finger protein). These four cathepsin-B copies are all part of a known derivation of proteins where the catalytic cysteine has been replaced by a serine residue (**Figure 40**) (Mendoza-Palomares et al. 2008). Despite this peculiarity, they are homologous to the cathepsin-B genes found in the core chromosomes, showing no evidence for a *T. congolense*-specific adaptation of cathepsin-B to the telomeres.

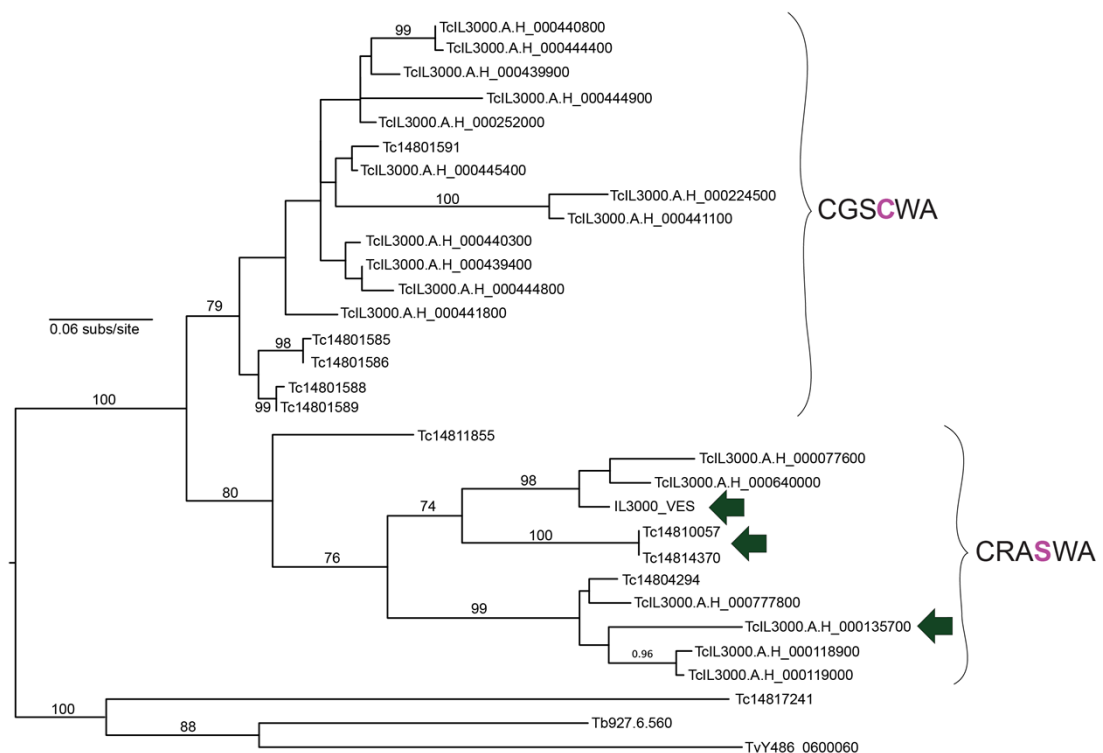


Figure 40 Consensus maximum likelihood phylogeny of cathepsin B protein sequences from African Trypanosomes. The phylogeny was estimated with PHYML with a

WAG+ Γ model and 100 bootstrap replicates. Bootstrap percentages higher than 70 % are shown in the internal nodes. Green arrows show the position of telomeric genes. The amino acid substitution cysteine to serine is shown in pink. Tree is mid-rooted.

DEAH-box RNA helicase

DEAH-box RNA helicases were found inside the ES in 17 telomere-containing contigs from Tc1/148 and 7 from IL3000. These genes were occasionally arranged in tandem pairs. Phylogenetic analysis showed that ES copies cluster with subtelomeric copies, but have likely derived from a specific helicase from chromosome 6 (**Figure 41**). This helicase exists in all trypanosomatids in a similar genomic context; its locus is characterised by a strand-switch region at the 5' end, followed by a conserved serine/threonine protein phosphatase that has been lost in *T. congolense*, only. At the 3' end, it is flanked by several conserved genes, including a dephospho-CoA kinase (TcIL3000_6_260).

The phylogeny suggests that *T. congolense* expanded this particular copy of DEAH-box RNA helicases from the core into the subtelomere. Often, these genes can be transposed to the telomeres, by unknown mechanisms. To test the contribution of selection to the expansion of this gene family, we searched for evidence of recombination using GARD (Pond et al. 2006) and subsequently performed three tests of site-level selection [FEL, FUBAR, and REL (Pond & Frost 2005; Murrell et al. 2013)] and one test of branch-level selection (BSR) on non-recombinant sequences. Two significant breakpoints were found at nucleotide 873 and 1522, but inspection of the sequence alignment did not reveal an obvious recombination point. Regarding the role of selection, all tests agreed on one site being under diversifying selection, whilst the remaining sites showed evidence for purifying selection, suggesting that the DEAH/box RNA helicase expansion in *T. congolense* is not driven by positive selection or gene conversion. Together, these results indicate that the DEAH-box RNA helicase expansion in *T. congolense* is not a telomeric-unique phenomenon.

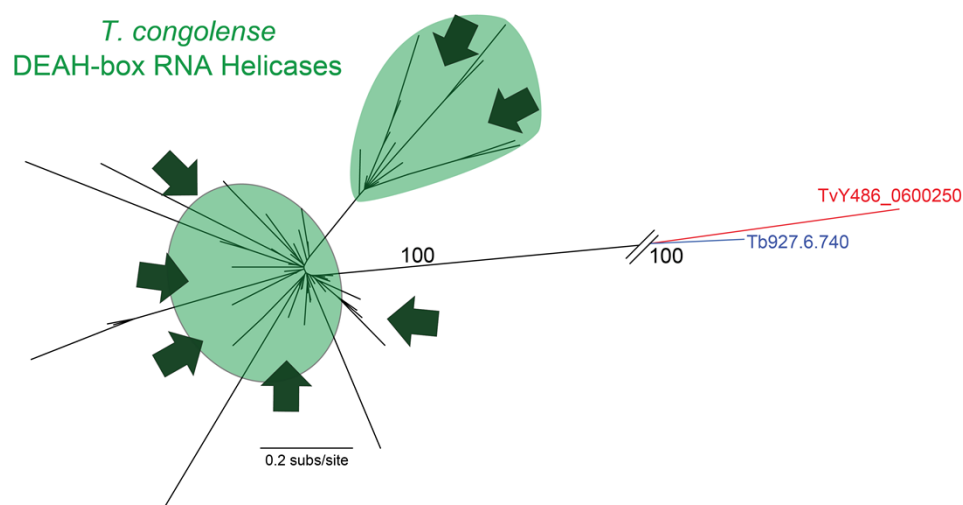


Figure 41 Consensus maximum likelihood phylogeny of DEAH-box RNA helicase protein sequences from *Trypanosoma congolense* and *Trypanosoma brucei*. The phylogeny was estimated with PHYML with a VT+ Γ +F model and 100 bootstrap replicates. Branches are labelled with bootstrap percentages higher than 70 %. Green arrows show the position of telomeric genes.

Retroposon hot spot protein pseudogenes were found inside the ES in fifteen contigs. One is found within or in the boundary of the ES in 6 contigs (i.e. TcIL3000_0_56340) and expression data suggest it is expressed in both metacyclic and bloodstream form life stages, and three seem to be preferentially expressed in the metacyclic stage (TcIL3000_0_32470, TcIL3000_0_32480, TcIL3000_0_32490) (Helm et al. 2009, The Wellcome Sanger Institute, ENA sample accession number SAMEA3629513). The latter belong to a subtelomeric tandem array and they exist in at least eight more instances throughout the subtelomeres. Due to their short length, it is likely that the expression data refers to the subtelomeric copy of the genes. Their role, if any, in VSG expression, remains unclear, but their absence from the IL3000 ES suggests they may result from sporadic, arbitrary translocations to the Tc1/148 ES.

Six other coding sequences were found inside the Tc1/148 ES. These corresponded to four predicted hypothetical proteins in the IL3000 genome (**Table 10**). Three genes have been shown to be expressed, either by EST library (Helm et al. 2009), metacyclic/epimastigote transcriptomes (chapter 3) or bloodstream form

transcriptome (The Wellcome Sanger Institute, ENA sample accession number SAMEA3629513).

Table 10 Hypothetical proteins found inside the ES of Tc1/148 and their available expression data. IDs are given for the IL3000 best hits obtained by sequence similarity search.

Hypothetical Protein	Contigs	Expression Data		
		<i>EST library</i>	<i>MF/EM Transcriptomes</i>	<i>BSF Transcriptome</i>
TcIL3000_0_16860	3	No	No	Yes
TcIL3000_0_59850	1	No	No	Yes
TcIL3000_0_02720	1	No	Yes	Yes
TcIL3000_0_51940	1	No	No	No

Together, these results indicate that whilst there are *T. congolense* ESAGs, they are either paralogous or analogous, but not orthologous to *T. brucei* ESAGs and none approach their exclusivity, as they are not functionally distinct isoforms found uniquely in the ES.

4.3.4 Tree Topology

To explore the role of recombination in the *T. congolense* telomere evolution, phylogenetic relationships between different CNRs were investigated by likelihood comparisons of optimal and constrained trees (**Figure 42**). In the absence of recombination, the phylogenetic relationships between features along the telomeric structure should be constant, only reflecting the pattern of telomere duplication. However, if homologous recombination plays a role in sequence evolution, as observed in *T. brucei* ES, the optimal topology for each feature will be distinct to reflect the pattern of sequence exchange and telomeric reorganisation.

The phylogenetic signal varies along the *T. congolense* ES as suggested by the visible differences in between the optimal phylogenies of the CNR1 and CNR2-4. These differences are evident at the topological level and the comparison of their log-likelihood scores reveals they are statistically significant ($p < 0.01$). For all four combinations, comparing the likelihood score of the optimal topology with the constrained topology score resulted in a decrease, although a linear relationship with physical distance could not be established. The highest topological dissimilarity occurs between the CNR1 and CNR2, suggesting that a recombination point might

exist between these loci. However, further analysis of recombination using GARD did not reveal any significant breakpoints within the conserved sequences.

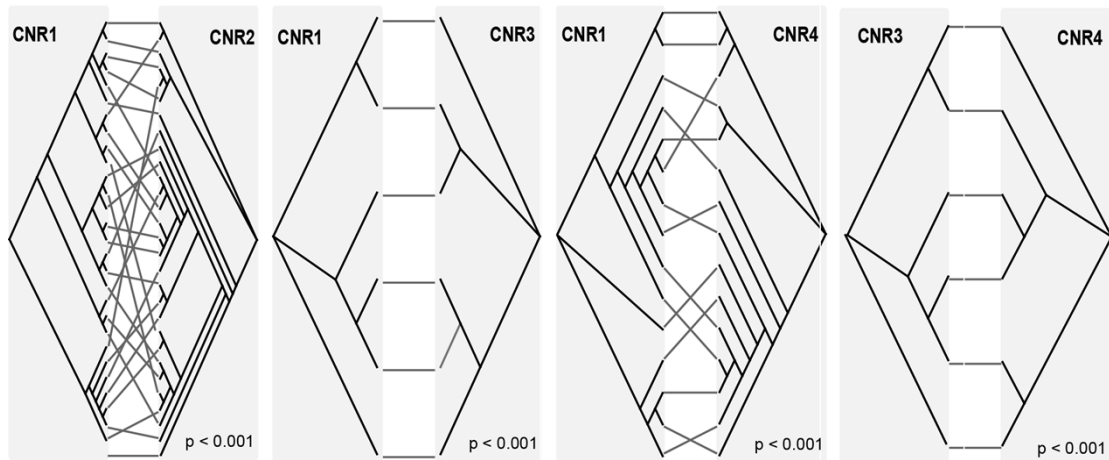


Figure 42 The difference in phylogenetic signal along the conserved non-coding regions of the ES. The first three tanglegrams relate the CNR1 phylogeny with those of CNR2-4 and the last relates CNR3 with CNR4. Black lines link corresponding contigs. Incongruence is highest in CNR2 and lowest in CNR3, although still significant.

4.3.5 The chromosomal location of the telomere-associated structures

From the Tc1/148 genome assembly, three contigs undoubtedly belong to the megabase chromosomes due to their length (1.78Mb, 1.57Mb, and 1.16Mb) and polycistronic gene organisation (**Figure 43**). The 1.78Mb contig belongs to chromosome 10 and does not contain the canonical structure described in this chapter. Instead, it is comprised of multiple VSGs and one partial copy of CNR3. At the subtelomeres, multiple copies of VSG and DEAH-box RNA helicase also exist. Additionally, at the core end of the 1.78 Mb contig, there is the only evidence of CNR4 outside the telomeric context. The 1.57 Mb contig belongs to chromosome 11 and contains a canonical structure: the 369 bp repeat, preceded by ingi; CNR1 in the same strand; one frameshifted VSG in the complementary DNA strand; and CNR4 immediately upstream of the telomere. In the core and subtelomere of the 1.57Mb contig, multiple VSG copies and one DEAH-box RNA helicase were found. The 1.16Mb contig belongs to chromosome 6 and also lacks a complete canonical

telomeric structure. However, it does have a VSG pseudogene and two adjacent copies of CNR3 and CNR4 in the leading DNA strand as the last features upstream the telomere. VSGs were not identified anywhere else in the contig, but two copies of DEAH-box RNA helicases, two copies of cathepsin-B, and one ESAG8-like gene were found dispersed along the subtelomere and core. In IL3000, contigs were not long enough to be undoubtedly assigned to the megabase chromosomes, although the scaffolded assembly has allocated two telomere-containing contigs to chromosomes 9 and 11, both of which contain the full canonical structure described here.

The Tc1/148 genome assembly also resulted in 25 complete mini-chromosomes with sizes ranging from 20,914 to 37,974 bp. The general structure of the mini-chromosomes includes a long string of the 369 bp repeat in the centre, flanked by CNR1, VSGs, CNR2-4 and other coding regions. **Figure 44** shows all the structural combinations found. Nineteen out of twenty structures have VSGs, always as the most telomere-proximal coding region. In two instances, the transferrin receptor is in this position. All but two of the full mini-chromosomes recovered have at least one CNR1, always adjacent to the 369 bp repeat. The CNR2 is present in 4 of the 20 topologies and is always located adjacent to the CNR1. Conversely, CNR3 and CNR4 were only found in 2 topologies, always after the VSG but immediately preceding the telomere. The IL3000 genome assembly resulted in 6 complete mini-chromosomes with sizes ranging from 21,075 to 37,994 bp. The general structure is identical to that described for Tc1/148 with the only exception that the 369 bp repeat is occasionally palindromic. These results indicate that the telomere-associated canonical structures are commonly present in the mini-chromosome ends.

In summary, this chapter presents more than one hundred telomere-associated structures in each strain (150 in Tc1/148 and 128 in IL3000). It shows that this is a canonical structure conserved across telomeres and strains, composed of one complex repeat, the VSG, and at least four conserved non-coding regions. Additionally, these structures may contain non-VSG coding regions of variable functions, including cathepsin B, DEAH-box RNA helicase, Fam53, and conserved hypothetical proteins. Finally, it suggests that the telomere-associated structures are subject to homologous recombination and sequence reorganisation. Together, these results show that, like *T. brucei*, *T. congolense* has canonical telomeric structures that are likely to be VSG ES. However, they are analogous and not homologous to the *T. brucei* ES.

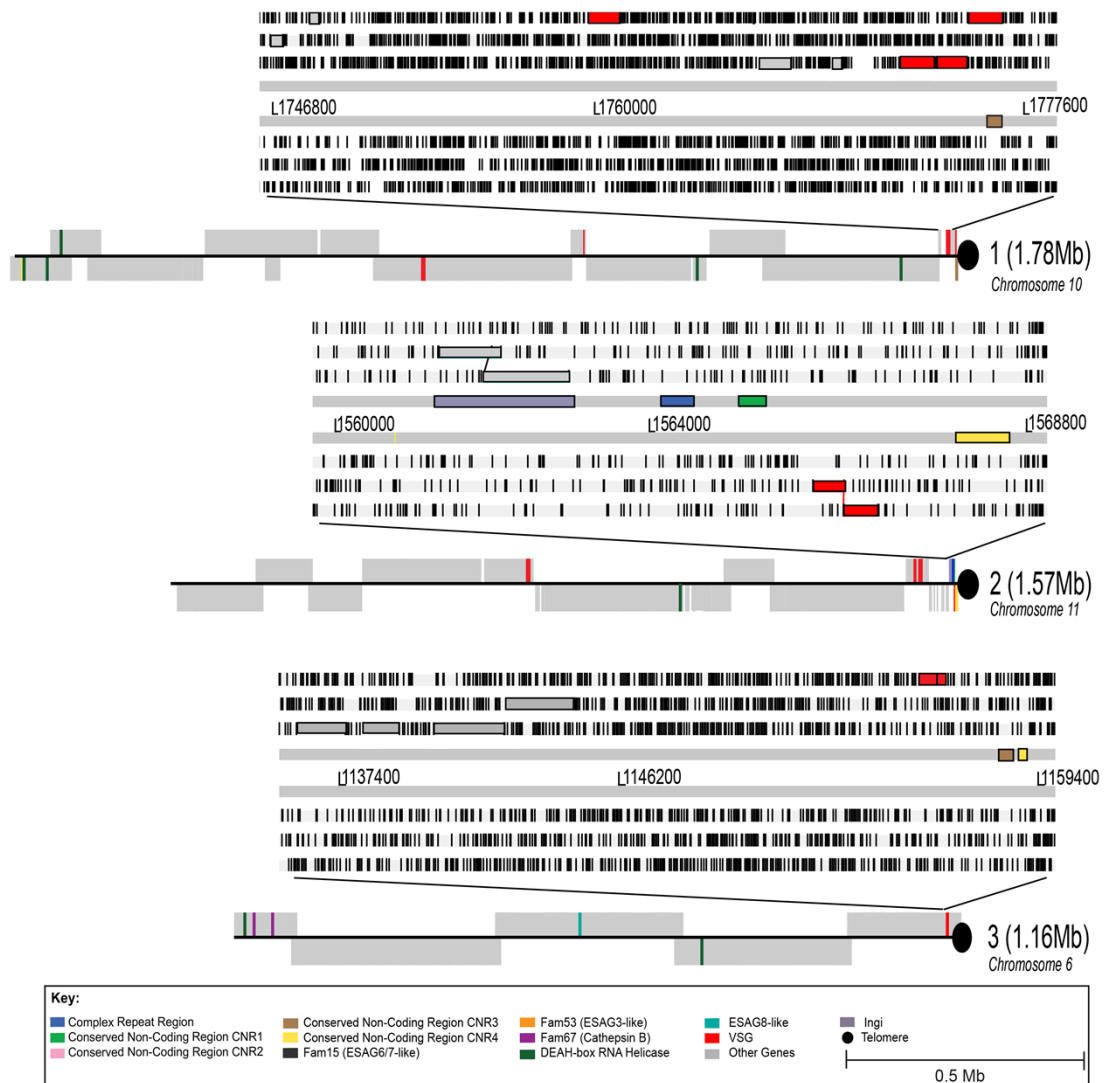


Figure 43 Structure of the telomere-containing contigs from the Tc1/148 megabase chromosomes. Contigs 1-3 belong to chromosome 10, 11, and 6, respectively, according to sequence similarity searches. Contigs are drawn to scale and have been aligned at their telomeric end. Conserved features are shown according to key.

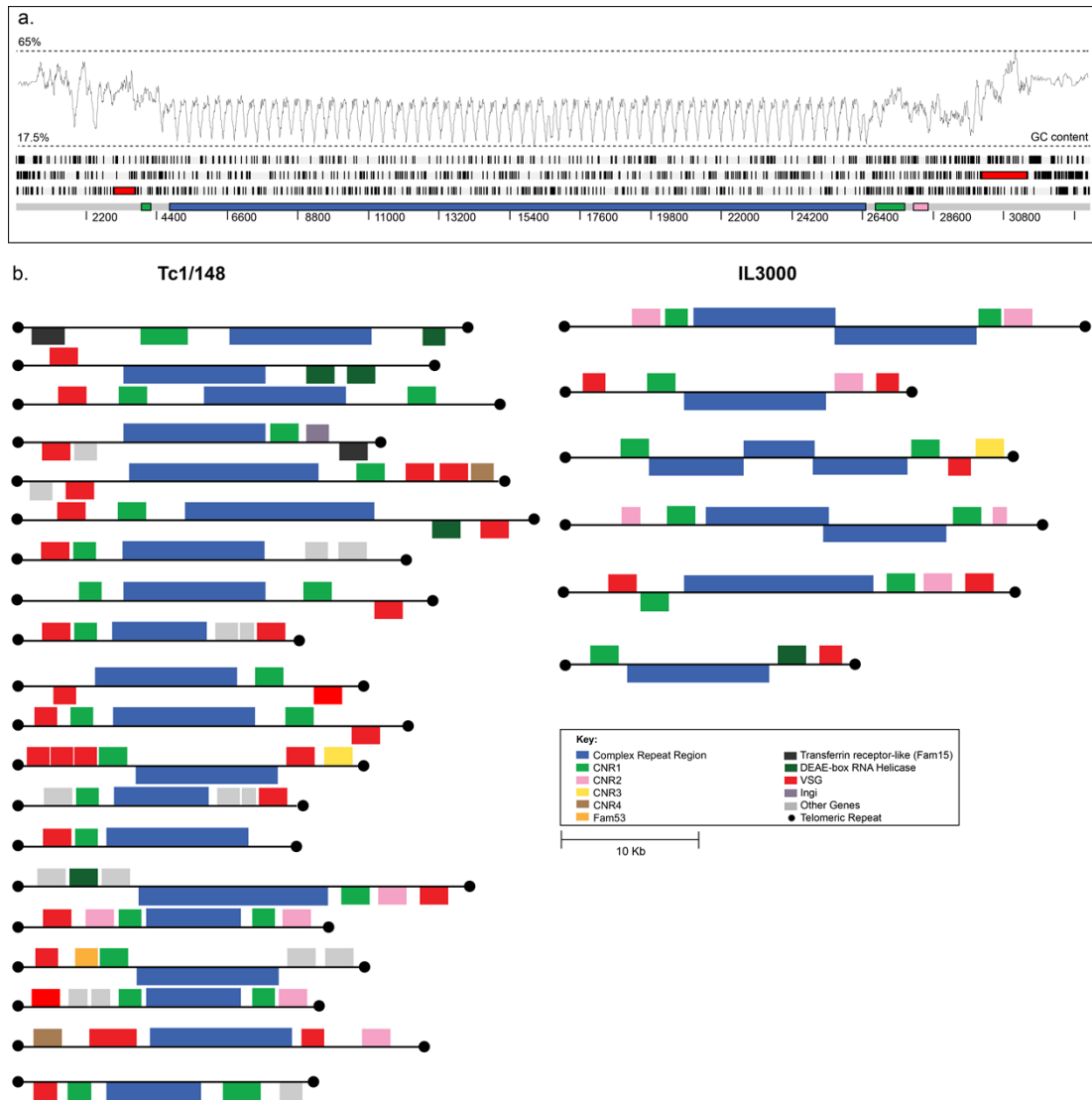


Figure 44 The non-redundant structure of the *T. congolense* mini-chromosomes in Tc1/148 and IL3000. A. Artemis plot of a complete mini-chromosome with a GC content graph. B. Mini-chromosomes are drawn to scale and have been aligned at their 5' end. Conserved features are shown according to key.

4.4 Discussion

This chapter identified and described VSG-rich structures at the telomeres of *T. congolense* using SMRT sequencing technology. The data presented shows that the *T. congolense* telomeres in multiple strains are associated with a repetitive domain, that has a canonical structure consisting of a complex repeat region, the VSG gene and four conserved non-coding regions, here named CNR1-4. In addition, they were shown to contain sporadic non-VSG genes, predominantly a specific lineage of DEAH-box RNA helicases, Fam53 and Fam15. These results were consistent in both Tc1/148 and IL3000.

Second-generation sequencing technologies (e.g. Illumina Sequencing) have been exceedingly useful, but they rely on short read lengths, which makes genome assembly difficult. This is particularly challenging for repetitive regions such as the telomeres and the mini-chromosomes, (Rhoads & Au 2015). For trypanosomes, short-read sequencing has resulted in genome assemblies containing large 'bins'. From an expression site point of view, these may potentially contain useful data. Long-read sequencing, such as SMRT sequencing (Pacific BioSciences, USA), may resolve this issue. Initially developed mainly for bacterial genomes, SMRT sequencing has improved to be applicable to large genomes. SMRT sequencing offers read lengths up to 60 kb, facilitating assembly of repetitive regions. As genomes get fragmented less often, they potentially conserve unique flanking sequences that can determine the origin of the read. The on-going improvements in the technology have also allowed for the polymerase read quality to be generally higher than 84 % and thus resolve the initial issue of higher error rate (Rhoads & Au 2015). Moreover, although SMRT cells are less efficient than short-read sequencing technologies both in terms of number of successful reads and number of reads yielded (e.g. $3.5\text{--}7.5 \times 10^4$ reads, compared to 1.2×10^9 paired reads per run in 250 bp Illumina HiSeq 2500 ; Rhoads & Au 2015), this issue can be partially overcome by using multiple SMRT cells. In this chapter, I have shown that SMRT sequencing can be helpful in describing the sequence diversity of telomeres and mini-chromosomes of African trypanosomes.

When the 369 bp repeat was discovered, it was proposed as a feature of the *T. congolense* mini-chromosomes (Kukla et al. 1987; Gibson et al. 1988; Moser et al. 1989), analogous to the 177 bp repeat in *T. brucei* (Sloof et al. 1983; Gibson et al.

1988), the 170 bp repeat in *T. vivax* (Dickin & Gibson 1989), the 550 bp repeat in *T. simiae* (Majiwa & Webster 1987) and the 195 bp repeat satellite DNA in *T. cruzi* (Gonzalez et al. 1984). All these repeats correspond to 5-10 % of the total nuclear genome and, with the exception of *T. vivax*, are AT-rich (29-35 % GC) (Gonzalez et al. 1984; Dickin & Gibson 1989; Moser et al. 1989). Indeed, the majority of contigs containing the 369 bp repeat described in this chapter belong to mini-chromosomes. However, I show that it can also be associated with megabase chromosomes. The conservation of the telomeric structures described in this study, the abundance of VSGs found, and the fact that the active VSG in IL3000 was found in a similar genomic context, suggest that these structures are *T. congolense* VSG expression sites. On this basis, a striking difference between *T. congolense* and *T. brucei* is that the majority of ES containing VSGs are found at the ends of mini-chromosomes.

T. brucei mini-chromosomes are rich in VSGs (Marcello & Barry 2007a; Cross et al. 2014). In *T. brucei*, mini-chromosomes are thought to be transcriptionally silent (Wickstead et al. 2003; Wickstead et al. 2004). The first instance of evidence supporting this conclusion comes from the observation that the 177 bp repeat is not transcribed (Sloof et al. 1983). Subsequently, members of the Cross and Borst laboratories showed that VSGs from mini-chromosomes must be converted to one of the ES in the megabase chromosomes to become active (de Lange et al. 1983; Van der Ploeg et al. 1984; Borst 1986; Cross 1990). More recent evidence corroborating the previous studies includes experiments of ectopic gene expression from the mini-chromosomes (Wickstead et al. 2002) and the transcriptomic analysis of *T. brucei* Lister 427, where very few reads map to sequences from the mini-chromosomal DNA (Siegel et al. 2010). Furthermore, the canonical features of the *T. brucei* ES, i.e. the ESAGs, are absent from the mini-chromosomes (Wickstead et al. 2004). In contrast, in *T. congolense* the same canonical structure is found throughout the mini-chromosomes and the megabase chromosomes. This might indicate that the ES of both chromosome types can become active. However, it could also be that all mini-chromosome and megabase chromosome ES captured in this analysis are silent, like they are in *T. brucei*.

Another difference in the mini-chromosomes between *T. congolense* and *T. brucei* is their number. The *T. brucei* TREU 927 genome contains approximately 100 mini-chromosomes of 30-150 kb and 5 intermediate chromosomes of 200-900 kb (Hertz-Fowler et al. 2007). In contrast, the Southern blot gel presented in this chapter (**Figure 33**) indicates that mini-chromosomes represent 7-10 Mb of DNA in *T.*

congolense IL3000, thus suggesting that a diploid *T. congolense* cell may have 240-320 individual mini-chromosomes (assuming an average mini-chromosome size of ~30 kb). Furthermore, the molecular karyotype of *T. congolense* has been shown to vary in different serodemes (Majiwa et al. 1986), a feature attributed to the aneuploidy of the mini-chromosomes and their non-Mendelian inheritance mode (Wells et al. 1987; Alsford et al. 2001). In this chapter, I describe more than 100 ES structures, all sharing a canonical structure and very low length and composition variation. The data suggest that *T. congolense* has more ES than *T. brucei*. How this relates to antigenic switching, VSG gene conversion from the subtelomeres to the ES, and to VSG sequence diversity generation remains unknown.

Structurally, the main differences between *T. congolense* and *T. brucei* ES are the nature and number of coding regions in the polycistronic transcription unit, and the existence of defined anchor points for sequence exchange. In *T. brucei*, ESAGs are a canonical feature of the BES (Becker et al. 2004; Hertz-Fowler et al. 2008). This species has thirteen characterised ESAGs, most of which are essential for activation and retain their order throughout the generic BES structure (Hertz-Fowler et al. 2008). With the exception of the facultative ESAG10, all *T. brucei* ESAGs were specifically recruited from the core and subtelomeres and independently diversified in the ES (Jackson et al. 2013). In contrast, *T. congolense* has few non-VSG coding regions and none showing the exclusivity and ES adaptation of *T. brucei* ESAGs. Although Fam15, Fam53, and cathepsin have been found in a small percentage of contigs, they lack the intrinsic features of an ESAG as defined here: i.e. they are indistinguishable from subtelomeric homologs, and found beyond the ES more often than within it. Therefore, these genes cannot represent a distinct telomere-associated adaptation in the manner of *T. brucei* ESAGs.

Among the non-VSG genes found in the ES, the DEAH-box RNA helicases could potentially be a *T. congolense* ESAG as it derives from a single-copy core chromosomal lineage. DEAH-box RNA helicases are remodelling enzymes that use ATP to catalyse the separation of double stranded RNA. They are conserved in most living organisms, including eukaryotes, bacteria and viruses (Marchat et al. 2015). In parasitic genomes, DEAH-box RNA helicases can be involved in many RNA-related processes, including pre-mRNA splicing in *Entamoeba histolytica* (Valdés et al. 2014); RNA silencing in trypanosomatids (Kramer et al. 2010; Holetz et al. 2010; Zinoviev et al. 2011), *Plasmodium falciparum* (Tarique et al. 2013), and *Giardia* (Adam 2000); and translation regulation in nematodes (Marchat et al. 2015).

However, the particular lineage of DEAH-box RNA helicases found in the *T. congolense* ES has not yet been functionally characterised.

When asking whether this helicase is an ESAG, it is worth noting that they are present in only in 5-11 % of the described ES and that they do not show any evidence for sequence divergence relative to their homologue (and presumed precursor) in the core genome. Although they have expanded in *T. congolense* from a single-copy core chromosomal lineage present in other trypanosomes, the sequences of the isoforms found in the ES are indistinguishable from the subtelomeric copies. This is unlike the specialization in sequence that we observe for the majority of *T. brucei* ESAGs.

T. congolense telomeric structure is consistent with the hypothesis that *T. brucei* ESAGs were independently recruited to the ES as *T. brucei*-specific innovations, and further suggests that *T. congolense* has not undergone this same process. Perhaps, this is because *T. congolense* has evolved different survival mechanisms that do not require them. The role of all ESAGs in *T. brucei* is not completely clear, but for example, *T. brucei* ESAG6 and 7, the transferrin receptors, are exclusive to the ES (with the exception of a single locus at a strand switch region) (Salmon et al. 1994; Schell et al. 1991; Jackson et al. 2013). Thus, all transcription of transferrin receptors must occur from the ES. In contrast, *T. congolense* has evolved a large repertoire of subtelomeric transferrin receptors (Jackson et al. 2013), therefore transcription from the ES may not be essential.

In terms of conserved non-coding regions, *T. brucei* has only two that are known: the promoter as the most upstream feature of the BES and the 70 bp repeat, which separates the VSG from the rest of the BES and often works as a recombination anchor point (Hertz-Fowler et al. 2008). Here I show that *T. congolense* has at least five conserved non-coding regions (the 369 bp repeat and CNR1-4). Though they do not exist in all telomeres, they maintain both the general order within the context and the relative distances to the VSG and the telomere. This suggests that CNR2-4 might play a role in gene regulation within the ES or, following the example in *T. brucei*, sequence transposition. The phylogenetic signal along the telomeric structure of *T. congolense* is not consistent. In fact, phylogenetic comparison of the cohort of CNR1 and CNR2-4 sequences shows differences in tree topology that is statistically significant, suggesting that these ES features are uncoupled through time, and indicating frequent sequence exchange between telomeres. Although the low number of telomeres containing all the conserved coding regions precludes an

exhaustive recombination analysis, the evidence suggests that, like *T. brucei*, the *T. congolense* ES also undergo frequent recombination.

Where and how recombination is happening remains unanswered. In *T. brucei*, quantifying the difference in phylogenetic signal along the BES revealed that sequence exchange occurs using various breakpoints, including at ESAG3, between ESAG5 and ESAG4/8, between ESAGs1/2 and 11, and at the 70 bp repeat (Hertz-Fowler et al. 2008). ESAG3 has been proposed as an important intervening feature in ES rearrangement, being the only ESAG appearing at multiple locations in the ES and not clustering by position (Hertz-Fowler et al. 2008). VSG switching is known to be facilitated by the 70 bp repeat upstream of the VSG and the VSG C-terminal domain (Hovel-Miner et al. 2016). These two structures act as recombination anchor points to enable VSG recruitment from the subtelomeres to the ES and VSG movement and conversion between ES (Marcello & Barry 2007a; Hovel-Miner et al. 2016). *T. congolense* VSGs lack homologous CTDs that would play such a role, having instead 15 distinct CTDs (Jackson et al. 2012). Furthermore, a structure analogous to the 70 bp repeat adjacent to *T. congolense* VSGs has not been found so far. It is tempting to speculate that the existing conserved non-coding regions might act as annealing regions to facilitate telomeric exchange and gene conversion between telomeres. CNR2 could work as the 5' recombination anchor point and CNR3/4 as the 3' recombination anchor point. However, as they are not found in the subtelomeres, it is unclear how the same regions might facilitate exchange with the vast majority of VSG loci.

If CNR2-4 are not involved in gene conversion and sequence transposition, they could be involved in the regulation of ES activation. In fact, non-coding regions are often associated with regulation of expression or chromosomal localisation of variant gene families. For instance, the *var* genes, which are linked to virulence and antigenic variation in *Plasmodium falciparum*, have a canonical structure in the subtelomeres and show a conserved gene organisation that is driven by their 5' noncoding sequences (Kyes et al. 2007). Moreover, their monoexpression is controlled by the variant silencing gene PfSETvs, which silences *var* gene transcription by methylating histone H3 Lysine 36. The methylated version of this histone coats not only the coding region of the *var* gene, but also its promoter, completely blocking transcription. To transcribe the active *var* gene, a long coding RNA, generated from the transcription initiation site in the antisense direction, inhibits PfSETvs, allowing transcription initiation (Jiang et al. 2013).

In eukaryotes, post-translational modifications of histones have been proposed to affect transcription activation, chromatin condensation and signalling for DNA repair (Strahl & Allis 2000). In *T. brucei* and *T. cruzi*, histone acetylation and methylation have indeed been described as mechanisms of epigenetic control (Figueiredo et al. 2008; Respuela et al. 2008; Daniell 2012). However, as gene transcription is mostly post-transcriptionally regulated, the repertoire of chromatin-modifying and chromatin-remodelling enzymes and the extent of histone post-translational modifications is considerably smaller than in most eukaryotes (reviewed in Figueiredo et al. 2009). Thus, only one histone-modifying enzyme (DOT1B) has been recognised to be directly involved in VSG transcriptional regulation and ES switching through trimethylation of H3K76 for rapid antigenic switching (Figueiredo et al. 2008).

In the bovine haemoparasite *Babesia bovis*, antigenic variation involves the variant erythrocyte surface antigen-1 (VESA1), which exists in quasi-palindromic loci comprised of a bidirectional promoter and non-coding regulatory regions flanking the variant antigens (Al-Khedery & Allred 2006; X. Wang et al. 2012). These non-coding sequences directly affect the levels of activity of the promoter and can drive expression from transcriptionally silent loci, suggesting their role in *in situ* transcriptional switching (X. Wang et al. 2012). Therefore, it would not be unprecedented for conserved non-coding regions to hold important functions in regulation of the VSG expression.

The VSGs found in the telomeres derive from most, but not all phylotypes. In particular, phylotype 6 and 9 were not observed in the telomeres of either Tc1/148 or IL3000. This is further emphasised by the observation from chapter 3 that phylotype 6 is barely expressed by metacyclic parasites. The low expression could be explained simply by developmental regulation, but its paucity in the ES of two distinct strains and the IL3000 bloodstream transcriptome suggest that these genes might have evolved a function unrelated to antigenic variation.

4.4.1 Future directions

The work presented in this chapter raises many important questions, but perhaps the most important ones for VSG biology are, first, to experimentally confirm that the structures I described here are VSG expression sites. Second is to understand

whether phylotype 6 and 9 encode functional VSG proteins, or if they have an invariant role. Identifying the active VSG through proteomics and identifying the genomic position of the cognate VSG gene can answer the first question, whilst ChIP-seq can be used to identify the expression site promoter by detecting its interaction with RNA polymerase I.

Discovering whether phylotype 6 and 9 encode non-VSG proteins could be achieved by following adapting the approach used for the characterisation of *T. brucei* VR genes by Marcello & Barry (2007). If they are not variant antigens, sequence homogenization might bring negative fitness consequences, therefore these genes are likely to exist in fewer numbers and be more conserved across the species. This hypothesis can be tested through the comparison of their repertoires in terms of sequence conservation, size, and phylogenetic relationships amongst strains. Particularly for phylotype 9, which is the smallest of the *T. congolense* VSG-like sequence repertoire, PCR primers could be designed to test their presence and sequence conservation across additional field isolates without the need for genome sequencing. Another aspect that could differentiate phylotypes 6 and 9 from the variant antigens is their expression level in the various parasite forms, which could be detected by RT-PCR in an inexpensive manner. When a similar experiment was conducted for the VR genes, the evidence for expression of VR genes by procyclic parasites strongly indicated their non-VSG role (Marcello & Barry 2007a), and the same is possible in this context. Beyond this, gene knockout studies and functional *in vitro* assays could elucidate their specific function.

Answering either question would require an investment in molecular biology studies applied to *T. congolense*. Molecular biology techniques have been mostly applied to *T. brucei*. The few studies performed in *T. congolense* have focused on establishing basic (but crucial) premises, such as PCR diagnosis, molecular karyotype characterisation, and cell culture establishment. Recently, more complex work is being performed, such as the characterisation of quorum-sensing signal responses in high-parasitaemia *T. congolense* infections and the inter-communication with *T. brucei* in mixed infections (Silvester et al. 2017). However, there is a need to establish models of infection and genetic engineering in *T. congolense* to fully understand the functional consequences of the VSG ES structure and characterise the VSG expression machinery in this parasite.

4.4.2 Conclusion

This chapter shows that *T. congolense* telomeric VSGs typically occur in a specific and conserved genomic context. These conserved structures are widely present in both strains analysed and are primarily associated with mini-chromosome telomeres, although we have also observed them on megabase chromosomes. These VSG-rich structures resemble *T. brucei* VSG ES in their location and conservation, but they lack abundant and conserved non-VSG coding regions (i.e. ESAGs). Instead, they are characterised by several conserved non-coding regions, and sporadic non-VSG genes, which do not show any unique sequence derivation specific to the telomeric environment. This indicates that *T. congolense* telomere-associated structures are analogous, rather than orthologous, to *T. brucei* ES. The most parsimonious explanation for such observation is that their common ancestor also had a telomeric ES structure, though any sequence homology has been lost through parasite divergence. The components of the ancestral structure will remain unknown, erased by large sequence diversification, but would likely have had an analogous expression site structure composed of a repeat, a promoter, the VSG, and the telomere.

Chapter 5. The Variant Antigen Profile to quantify antigen diversity in *Trypanosoma vivax*

5.1 Introduction

Trypanosoma vivax is an exclusively animal pathogen with a dixenic life cycle. Unlike *T. congolense* and *T. brucei*, it lacks a procyclic stage, colonising the proventriculus and foregut in the first day of infection and then migrating to the mouthparts (Ooi et al. 2016). Although *Glossina sp.* is the only vector species where *T. vivax* can grow and multiply, mechanical transmission is possible between other hematophagous flies, which has led to its introduction into South America, likely from West Africa (Osório et al. 2008). *T. vivax* transmission is solely mechanical in tsetse-free areas, but both cyclical and mechanical in Sub-Saharan Africa (Osório et al. 2008). In both areas, *T. vivax* can be mechanically transmitted by tabanids (*Cryptotylus* and *Tabanus* spp.) and *Stomoxys* spp. (Hoare 1972; Levine 1973). Since *T. vivax* cannot complete its life cycle outside the tsetse belt, South American *T. vivax* isolates remain genetically very close to West African isolates, whilst a large genetic distance separates them both from East African parasites (Osório et al. 2008).

T. vivax is the least studied of the three African trypanosomes species mainly due to the inability to maintain the different life stages of the parasite in culture until very recently (D'Archivio et al. 2011). However, major differences in gene expression and VSG evolution in this species (Jackson et al. 2012; Jackson et al. 2015), and its presence in South America, make it a critical object of research in the trypanosomiasis field. One reason why I want to know more about *T. vivax* VSG is the phenotypic variability of *T. vivax* infections. In West Africa *T. vivax* often causes an acute, rapidly fatal disease. In East Africa, most disease is mild and chronic, but sporadically it can occasionally result in haemorrhagic syndrome (Welde et al. 1983). As VSG play an essential role in host-parasite interactions, they may be involved in clinical outcome.

A surface coat of 12-15 nm is visible by electron microscopy in all African trypanosome bloodstream forms (Vickerman 1969). However, in *T. vivax* this coat appears to be thinner, suggesting a less abundant VSG coat (Vickerman 1969). Indeed, VSG comprise only 56 % of the total gene transcripts (Greif et al. 2013), as opposed to 98 % in *T. brucei*. The transcriptome analysis of the *T. vivax* isolates IL1392 further shows that the parasite surface contains several non-VSG, species-specific proteins (Jackson et al. 2015). So far, 23 parasite-specific gene families encoding putative cell-surface proteins have been described (Jackson et al. 2013). Whilst they have not been experimentally shown to localise at the surface, they encode the necessary diagnostic motifs (e.g. signal peptide, GPI-anchor) and they are abundantly expressed in the bloodstream stage of the parasite (Jackson et al. 2013; Jackson et al. 2015).

The VSG-like gene repertoire in *T. vivax* is comprised of four structurally different families (Fam23-26), of which Fam23 and 24 are homologous to *T. brucei* a- and b-type VSG, respectively (Jackson et al. 2012). Whether Fam25 and 26 encode functional variant antigens or if they have acquired alternative functions remains unresolved. As discussed in the previous chapters, functional differentiation of VSG-like sequences is well documented in both *T. brucei* and *T. congolense* (e.g. the transferrin receptors of *T. brucei* and *T. congolense* (Jackson et al. 2012); the SRA gene in *T. brucei rhodesiense* (De Greef & Hamers 1994; Van Xong et al. 1998); and the TgsGP gene in *T. brucei gambiense* (Uzureau et al. 2013; Capewell et al. 2013). So far, there is no evidence for the expression of Fam25/26 as variant antigens. In the study of *T. vivax* global gene expression, the identified superabundant BSF VSG belonged to Fam24 and two Fam25 members were preferentially expressed in the epimastigote stage, where VSG expression is repressed in other species (Jackson et al. 2015). Whilst the function of Fam25 and Fam26 remains undetermined, their role as variant antigens would be established if they were shown to provide the major surface glycoprotein in bloodstream infections.

At publication, the TvY486 reference genome included 721 VSGs. However, in the on-going re-sequencing project this has increased to 1920 (unpublished data produced by the Darby lab at the Institute of Integrative Biology of the University of Liverpool). In both cases, the relative sizes of each VSG-like subfamily remain constant. Fam23 is the largest, accounting for 50 % of the total repertoire, followed by Fam24 and Fam25, which account for 22 % and 18 % of the VSG pool,

respectively. Fam26 is the smallest family, contributing with 9 % of the genes. The phylogenetic distances between gene clades within each family are even more remarkable than in *T. congolense*, suggesting scarce recombination and explaining the why it has the lowest percentage of pseudogenes in African trypanosomes (15.5 % and 27.2 % of Fam23 and Fam24, respectively, compared to 21.1 % and 29.7 % of Fam13 and Fam16 in *T. congolense*, and 69.2 % and 72.2 % of a- and b-type VSG in *T. brucei*) (Berriman et al. 2005; Jackson et al. 2012). However, unlike *T. congolense*, intra-phylotype recombination is thought to be infrequent (Jackson et al. 2012).

The reference strain Y486, firstly isolated in Zaria, Nigeria, is a representative of West African and most South American *T. vivax* strains (Gibson 2012), but they are genetically distinct from East African strains (Cortez et al. 2006; Rodrigues et al. 2008; Adams et al. 2010). This may relate to the distinct pathological characteristics of East and West African *T. vivax* trypanosomiasis (Stephen 1986).

Overall, research into *T. vivax* molecular diversity supports the identification of East African *T. vivax* as a separate subspecies to West African and South American strains, mainly due to its distinctiveness in molecular diagnostic tests based on single molecular markers, e.g. *T. vivax*-specific antigen, SSU rRNA, and gGAPDH genes (Masake et al. 1997; Morlais et al. 2001; Njiru et al. 2004; Malele et al. 2003; Ventura et al. 2001; Adams et al. 2010). Furthermore, these studies also indicate greater diversity in East Africa compared to West Africa and South America (Rodrigues et al. 2017). However, considering that all of these studies were conducted with few isolates, they are not definitive, especially given that divergences at the SSU rRNA sequences amongst *T. congolense* subgroups are greater than those of the *Duttonella* subgenus (Rodrigues et al. 2008).

At first glance, I would expect the *T. vivax* VSG repertoire to be distinct between West and East African isolates, reflecting their genomic signature. However, given the ancestral origin of VSG lineages (Jackson et al. 2012), the evidence for low recombination within the *T. vivax* VSG repertoire, and the fact that the VSG repertoires of *T. congolense* 'savannah' and forest sub-types comprise identical VSG phylotypes (as observed in chapter 2), I would predict that the *T. vivax* VSG repertoire is conserved across the species.

Past research reports cases of *T. vivax* infections in cattle with undetectable parasitaemia and symptom recovery after 100 days post-infection (Barry 1986; Fidelis Jr et al. 2016). This suggests that protective immunity can occur naturally, perhaps a sign of antigenic exhaustion (Barry 1986). Research also shows that the VSGs expressed in early infection often belong to the same serodeme, depending on a combination of host species and size and pathogenicity of the parasite isolate, which suggest that VSG expression in *T. vivax* may be partially predictable (Barry 1986). To appreciate the scope of VSG diversity and recognise relationships between particular VSGs or VSG combinations and disease course can help to understand why *T. vivax* infections have major differences in morphology, host susceptibility, virulence, and pathology (Hoare 1972; Stephen 1986; Murray & Clarkson 1982).

The development of a VAP methodology universal for the *T. vivax* group would not only aid epidemiological mapping of AT, but also contribute to diagnostics and taxonomy by providing the largest comparison of continental-wide *T. vivax* isolates to date. As introduced in Chapter 2, the ultimate aim of the VAP is to be applicable to any organism employing antigenic variation for immune evasion, being sensitive enough to dissect diversity to the sequence level. Following the method development for *T. congolense* VSG, the first part of this chapter attempts to develop a methodology to characterise VSG diversity in *T. vivax* using a historical collection of 19 field strains.

In the second part, the methodology is applied to TvLins, a Brazilian strain of *T. vivax* isolated from an epidemic in dairy cows in the TvLins municipality of São Paulo, Brazil (Cadioli et al. 2012). South American *T. vivax* has expanded via mechanical transmission in non-tsetse vectors. All South American strains reported so far were closely related the reference TvY486, firstly isolated in Nigeria, West Africa (Cortez et al. 2006; Rodrigues et al. 2008). Whilst the majority of *T. vivax* infections in South America remain asymptomatic in endemic areas, when *T. vivax* escapes to non-endemic regions, it causes very acute disease with severe symptoms and rapid death. Likewise, TvLins infections caused acute disease with general symptoms, such as fever, jaundice, decreased milk production, weight loss, profuse diarrhoea, abortion, anaemia, leucocytosis, elevated plasma fibrinogen, and neurological symptoms, such as dysmetria, ataxia, muscle weakness, ptialism, lymph node enlargement and submandibular oedema (Cadioli et al. 2012). Despite the reasons behind such virulence remaining unknown, the parasite coat is of

particular interest due to its key role in host: parasite interactions. By applying an automated VAP methodology on genomic and transcriptomic reads, this chapter presents a detailed description of the TvLins VSG repertoire.

Specifically, this chapter aims to:

1. Investigate the universality of the VSG phylotypes identified in the reference strain using phylogenetic analysis;
2. Present a novel method to dissect antigenic diversity in *T. vivax* based on presence and absence of diagnostic VSGs using a collection of 19 historical field isolates;
3. Apply the method to quantify and describe antigenic diversity of the TvLins isolate in the context of the collection of historical isolates.

5.2 Methods

5.2.1 Sample identification and storage

Samples were collected between 1966 and 1990 by ILRI staff from cattle and tsetse flies as part of various experiments (**Table 11**). They were passaged into a naïve host (rodent, goat, or bovine) to increase parasitaemia. At peak parasitaemia, animals were bled and multiple 35 to 50 µl blood stabilates in 10-30 % glycerol were prepared and stored at -80 °C. Samples were subsequently transferred to liquid nitrogen. Samples for this project were chosen from the ILRI Biobank (Kenya) based on availability of at least 150µl of blood and number of passages. All samples used had had a maximum of 4 passages, as passaging through multiple hosts may affect the VSG repertoire. DNA was extracted from the blood stabilates in ILRI (Kenya) in January 2016 and imported immediately to the United Kingdom on dry ice.

The Brazilian sample *T. vivax* Lins used as an example application of the VAP method was first isolated from a dairy cow in the municipality of Lins, São Paulo, Brazil (Cadioli et al. 2012). Parasites were recovered from experimental infections of goats (*Capra aegagrus*) in the Federal University of São Paulo, Brazil. Two goats were experimentally infected with 1.25×10^5 parasites collected from a naturally infected cattle and cryopreserved in 8 % glycerol and 10 % EDTA. Parasitaemia was determined daily by haematocrit centrifugation technique as described by Woo (1970). When parasitaemia reached $\sim 3 \times 10^6$ trypanosomes/ml, 100 ml of goat blood was collected from the jugular vein into tubes without anticoagulant. After blood clotting, the blood was left at 27 °C for 1 hour to allow for parasites to move from the buffy coat into the serum. The serum with trypanosomes was collected into 1.5 centrifuge tubes (Eppendorf, USA) and centrifuged at 10,000 g for 30 min at 10°C. The pellet was mixed with 1 ml of Percoll (Sigma) containing 8.55 % sucrose, 2.0 % glucose, and the pH was adjusted to 7.4. The mixture was centrifuged at 17,500g for 20 min at 4 °C. Parasites were recovered from both the top and the middle layer of the Percoll gradient and resuspended 1:3 with PBS containing 0.5 % glucose (sodium phosphate 40 mM, pH 7.5, NaCl 150 mM) (PBSG). The solution was centrifuged at 4,500 g for 30 min at 4 °C to recover a parasite pellet. The pellet was washed twice in 40 ml of PBS with glucose to remove residual Percoll. The parasites were resuspended in 2 ml of PBS with glucose and used for DNA or RNA extraction.

Table 11 Sample ID, date and location of collection, host, and species used for passaging.

ID	Date	Location	Host	Passage species
ILV-21	1972	Antapar Teso, Uganda	Bovine	Goat
IL3658	1990	Ivory Coast	Bovine	unknown
IL3638	1990	Ivory Coast	Bovine	unknown
IL3651	1990	Ivory Coast	Bovine	Rat
IL2005	1969	Lugala, Uganda	Tsetse Fly	Goat
IL2714	1969	Lugala, Uganda	Tsetse Fly	Rat
IL2323	1969	Luuka, Uganda	Tsetse Fly	Rat
IL3171	N/A	The Gambia	Bovine	Bovine
IL684	1973	Yakawada, Nigeria	Bovine	Mice
IL1392	1981	Yakawada, Nigeria	Bovine	Goat
IL596	1973	Yakawada, Nigeria	Bovine	Mice
IL493	1973	Yakwada, Nigeria	Bovine	Mice
IL462	1973	Yakwada, Nigeria	Bovine	Mice
IL465	1973	Yakwada, Nigeria	Bovine	Mice
IL338	1973	Yakwada, Nigeria	Bovine	Mice
IL340	1962	Zaria, Nigeria	Bovine	Mice
IL11	1973	Zaria, Nigeria	Bovine	Mice
IL306	1973	Zaria, Nigeria	Bovine	Mice
IL319	1973	Zaria, Nigeria	Bovine	Mice
TvLins	2012	São Paulo, Brazil	Bovine	Goat

5.2.2 Cell lysis and nucleic acid extraction

DNA was extracted from the blood using a magnetic bead protocol adjusted to increase recovery of fractionated DNA. 200 µl blood stabilates were transferred to 2.0 ml Lo-Bind tubes (Eppendorf, UK). 1.5 ml ACK lysing buffer (0.15 M NH₄Cl, 10 mM KHCO₃, 0.1 mM EDTA) were added and the solution was incubated 3 min at room temperature, followed by a 10-min centrifugation at 650 g. The supernatant was discarded and the pellet washed in 500 µl MACS buffer (2 mM EDTA, 5xBSA in PBS pH7.2). The samples were centrifuged for 10 min at 650 g. The supernatant was discarded and cells were washed a second time in 750 µl MACS buffer and centrifuged for 15 min at 650 g. The supernatant was discarded and pellet resuspended in 100 µl lysis buffer (aqueous solution of 1 M Tris-HCl pH8.0, 0.1 mM NaCl, 10 µM EDTA, 5 % SDS, 0.14 µM Proteinase K). Samples were incubated at room temperature for 1 hour and DNA was extracted with the magnetic Sera-Mag Speedbeads protocol used for *T. congolense*. Recovered DNA was quantified with

Qubit® fluorometric dsDNA quantitation (dsDNA HS Assay Kit) (Life Technologies, UK) and sent for sequencing.

For TvLins, DNA was extracted with the phenol-chloroform method by Brazilian collaborators. One parasite sample for each infected goat was used to extract RNA. The two RNA replicates (S3 and S4) were obtained using the Dynabeads® mRNA DIRECT™ Kit (Thermo Fisher Scientific) according to the manufacturer's protocol.

5.2.3 Next-Generation Sequencing (NGS)

Genomic libraries were prepared at the Wellcome Trust Sanger Institute (Cambridgeshire, UK) using the NEBNext® Ultra™ DNA Library Prep Kit according to the manufacturer's protocol (New England Biolabs, UK). The protocol consists of mechanical shearing of DNA in 500 bp fragments using the Covaris ADA process (Covaris, UK), enzymatic DNA end repair and adapter ligation, size selection using magnetic Sera-Mag Speedbeads, and PCR Enrichment of Adaptor Ligated DNA. Genomic libraries were sequenced multiplexed in one lane over ten runs in a MiSeq platform (Illumina Inc, USA) as 150 bp paired-end (PE) reads.

For TvLins, gDNA was used to prepare 550 bp insert libraries using the TruSeq Nano DNA LT Library Preparation Kit, and sequenced on an Illumina MiSeq platform with paired-end reads (2 x 300 bp) obtained using a MiSeq Reagent Kits v3 (600 cycles). RNA samples were used to prepare a double-stranded cDNA library using the TruSeq Stranded mRNA LT Sample Preparation Kit (Illumina, San Diego, CA, USA) and sequenced on an Illumina MiSeq platform with paired-end reads (2x75 bp) obtained using a MiSeq Reagent Kits v3 (150 cycles).

Analysis of NGS data

Paired reads from the host species were removed by mapping all reads to the host genome (version 9.0) and keeping only unmapped reads for downstream analysis, using Bowtie2 under sensitive settings. Recovered reads were used to produce *de novo* assemblies of each strain.

Recovered reads were also mapped against the *T. vivax* Y486 reference genome. Mapping results were used to calculate per base coverage of sequenced strains

against the reference genome. Per base coverage was calculated from a sorted bam file using the genomecov option of the bedtools package. Total coverage was calculated with an in-house script counting the percentage of bases with sequencing coverage of at least 1.

To check that all samples derived from single infections, allele frequencies were calculated with vcftools. This allowed inference of haplotypes and detection of fluctuating allele frequencies inconsistent with diploidy. All frequencies from all samples were 0, 0.5, or 1, thus reflecting single infections.

De novo assembly

Recovered reads were used to produce *de novo* assemblies of each strain using velvet version 1.2.07 (Zerbino & Birney 2008). For the historical isolates, the following settings were applied: kmer of 65, insert length adjusted for 400 base pairs with standard deviation of 50, minimum pair count of 20, and coverage cutoff between 0 and 5, depending on sample quality. For TvLins the kmer was 99, producing an assembly of 178,079 contigs with n50 of 1,837.

The methodological process from contig assembly to VAP development is shown in **Figure 45** and consists of three main steps: identification of VSG-like sequences, VSG phylogeny analysis, and VAP development.

5.2.4 VSG-like sequence recovery

Assembled contigs were examined for VSG-like sequences by sequence similarity search with tBLASTx using a database of *T. vivax* Y486 VSG genes as query (**Figure 45**). A threshold of p-value > 0.001, contig length > 150 amino acids, and % identity \geq 75 was applied to select significant results. Sequences with 40 to 75 % similarity to the reference were manually inspected and its inclusion in the analysis empirically decided. The closest relative of each sequence query retrieved was used to assign recovered VSG-like sequences to a family and phylogeny.

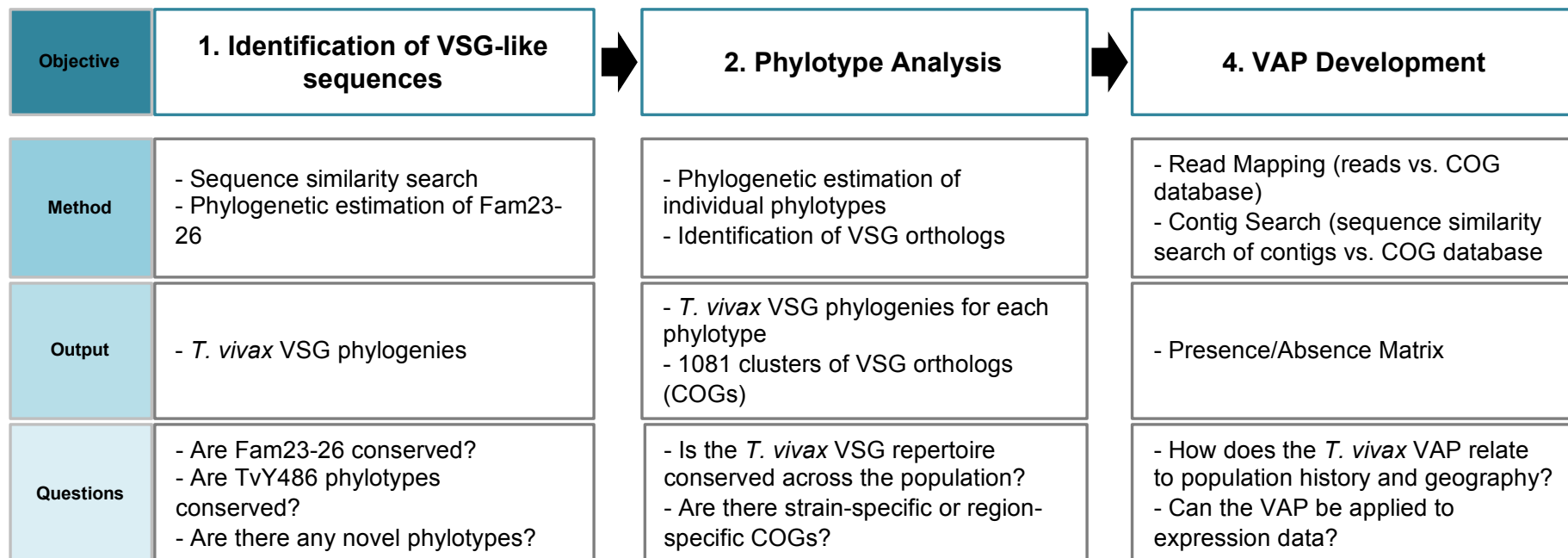


Figure 45 Methodology followed for the *T. vivax* VAP development. VSG-like sequences were recovered from the assembled contigs of each strain by sequence similarity search (tBLASTx). Translated nucleotide sequences were separated into Fam23 to Fam26, joined to the TvY486 VSG sequences and used to estimate 4 *T. vivax* VSG phylogenies. After each phylogeny was analysed, I concluded that phylotypes were widespread, but that universality could not be inferred due to missing data. Therefore, individual phylotypes were separately analysed to identify VSG orthologues that could be used as diagnostic sequences in the VAP. This resulted in 1081 COGs that were used to develop the VAP. The VAP methodology was developed using two different approaches: the first by mapping genomic reads to the COG database; the second by performing sequence similarity searches of the COG database in assembled contigs. Both approaches resulted in a presence/absence binary matrix that identifies location-specific COGs and provides the VSG repertoire relationships among strains.

5.2.5 Multiple Sequence Alignment

To investigate VSG family structure and phylotype universality, VSG-like sequences were aligned by family into four multiple sequence alignments (**Figure 45**). VSG-like nucleotide sequences were manually retrieved from the assembled contigs files and translated with BioEdit 7.2.5 (<http://www.mbio.ncsu.edu/BioEdit/bioedit.html>). Translated nucleotide sequences were aligned with ClustalW (Larkin et al. 2007), producing a nucleotide alignment for each of 96 phlotypes. Phlotypes belonging to the same family in the reference were manually joined and curated to produce an amino acid alignment for each family after partial sequence removal. The Fam23 alignment contained 1223 protein sequences of 120 to 272 amino acids; the Fam24 alignment contained 454 protein sequences of 110 to 336 amino acids, the Fam25 alignment contained 388 protein sequences of 149 to 325 amino acids, and the Fam26 alignment contained 165 protein sequences of 115 to 196 amino acids.

5.2.6 Phylogenetic estimation

VSG-like gene phylogenies were inspected to check whether all sequences clustered with co-orthologues from the reference genome and to detect new VSG clades not present in the reference genome (**Figure 45**). VSG phylogenies were estimated for each family from protein sequence alignments of recovered VSG-like sequences and TvY486 VSG sequences. Four ML phylogenies were built with the WAG+I^Γ substitution model (Whelan and Goldman 2001) after model selection in MEGA7 (Kumar et al. 2016). Robustness was assessed with 100 bootstrap replicates, using RaxML (Stamatakis 2014). Bayesian inference failed to converge on a consensus phylogeny for all four families.

To dissect further the phylogenetic structure of each phylotype, I searched for VSG co-orthologues (**Figure 45**). ML VSG phylogenies of each phylotype containing more than 4 sequences were estimated from a translated nucleotide sequence alignment of recovered VSG-like sequences and Y486 VSG sequences using the Tamura-Nei substitution model (Tamura & Nei 1993) in MEGA7 (Kumar et al. 2016).

5.2.7 Clusters of Orthologous Groups (COGs) identification

Each phylotype phylogeny was inspected to separate each member of each phylotype into defined groups based on their sequence identity (**Figure 45**). In the context of this chapter, VSG sequences of the same phylotype with more than 98 % nucleotide identity were considered co-orthologous and analysed as a single gene. VSG sequences with no orthologues in the sample set were included as individual orthologue groups. This approach resulted in a non-redundant gene database that allowed screening of genomes for specific VSG. In total, 1081 COGs were identified.

5.2.8 Binary matrices

Binary matrices were produced based on presence or absence of COGs using read mapping or contig similarity search (**Figure 45**). Mapping was performed using bowtie2 version 2.2.5 (Langmead & Salzberg 2012) under very sensitive and sensitive settings. COGs with at least one read mapped were considered to be present in the strain. Similarity search was performed with BLASTn (Altschul et al. 1990) with a 95 or 98 % identity, 0.001 E-value and 100 nucleotide length thresholds. The search and processing of the output to produce the binary matrix was done with an in-house script.

5.2.9 Strain variation

MiSeq reads were retrieved and mapped against the *T. vivax* Y486 genome using BWA mem (Li 2013), converted to bam format, and sorted and indexed with Samtools (Li et al. 2009). Sorted bam files were cleaned, and duplicates marked and indexed with Picard (<http://broadinstitute.github.io/picard/>). SNPs were called and filtered with Genome Analysis Toolkit suite according to the recommended protocol for multi-sample variant calling (Van der Auwera et al. 2013). Reads were realigned and loci called with GATK (Van der Auwera et al. 2013). SNPs were called for individual samples with HaplotypeCaller, and then the genotypes were called simultaneously for all samples. Finally, SNPs were extracted as a multi-sample file and filtered using default parameters using GATK (Van der Auwera et al. 2013). The multi-sample vcf file obtained from GATK was converted to fasta format using vcftools v0.1.14 (Danecek et al. 2011) and a maximum likelihood phylogeny was

estimated with RAxML, using the JTT+ Γ model of nucleotide substitution, following nucleotide model selection on MEGA7 (Kumar et al. 2016).

5.2.10 *T. vivax* Lins transcriptome analysis

TvLins transcriptomic reads were mapped to the reference TvY486 with Bowtie2 version 2.2.5 (Langmead & Salzberg 2012) under default settings. Mapping percentages of 90.35 % and 89.24 % were obtained for S3 and S4, respectively. Transcript abundances for both replicates were estimated using cufflinks version 2.2.1 (Trapnell et al. 2010), under default settings.

5.2.11 Ethics Statement

The animal work performed to obtain the genome and transcriptomes of TvLins was performed in accordance with the guidelines of the Brazilian College of Animal Experimentation (CONCEA), following the Brazilian law for “Procedures for the Scientific Use of Animals” (11.794/ 2008 and decree 6.899/2009). Ethical approval was obtained from the Ethical Committee to the Use of Animals (CEUA) of the Veterinary and Agrarian Sciences Faculty (FCAV) of the State University of São Paulo (Jaboticabal campus) (São Paulo, Brazil).

5.3 Results

In this chapter I present VAPs of 19 *T. vivax* isolates. Through the analysis of these VSG repertoires, I have investigated the level of conservation of Fam23-26 across the *T. vivax* species and I have dissected the TvY486 VSG repertoire into 96 phylotypes. I have investigated whether these phylotypes were universal, whether there are novel phylotypes, and whether they could be used to develop a *T. vivax* VAP methodology. The results show that a *T. vivax*-specific methodology is necessary, based on VSG orthologues rather than individual phylotypes. In the second part of this chapter, I apply the VAP to the analysis of the TvLins transcriptome to show that the VAP can be used to measure and compare VSG expression across infections.

5.3.1 Genome completion of sequenced field strains

Nineteen *T. vivax* isolates from four African countries were successfully sequenced (**Table 11**). To understand how deep the sequencing process was, the depth of coverage per base for each isolate was calculated (**Table 12**). This metric shows the completeness of the sequencing data, and therefore estimates data reliability. The majority of samples (15/19) had a genome coverage higher than 70 %; four samples, Tv319, Tv2005, TvILV-21 and Tv3171, were 51 % to 70 % complete. These were still included in the analysis, but flagged.

Table 12 Sequencing coverage and mapping results of field strains compared to the reference strain. Genome coverage was calculated as number of reference nucleotide bases present in at least one read over the total number of nucleotide bases. Mapping proportion was calculated as the proportion of paired reads mapping to the reference genome with Bowtie2 (Langmead & Salzberg 2012).

Strain	Genome coverage	Mapping to <i>T. vivax</i>
Tv11	0.79	0.39
Tv1392	0.77	0.79
Tv2005	0.61	0.04
Tv2323	0.71	0.53
Tv2714	0.70	0.59
Tv306	0.79	0.31
Tv3171	0.55	0.02
Tv319	0.70	0.11
Tv338	0.80	0.11
Tv340	0.78	0.24
Tv3638	0.72	0.30
Tv3651	0.71	0.25
Tv3658	0.72	0.20
Tv462	0.79	0.31
Tv465	0.79	0.28
Tv493	0.79	0.10
Tv596	0.80	0.09
Tv684	0.80	0.14
TvILV-21	0.51	0.03

5.3.2 Sampling test

Another metric to assess quality of data, in the context of this study, is the size and composition of the VSG repertoire that can be recovered by sequence similarity searches, using the previously defined length and identity thresholds (≥ 150 amino acids and ≥ 40 % ID). The number of VSGs recovered in the field strains ranged from 40 to 303, representing 6 to 42 % of the reference VSG genomic repertoire (**Table 13**). The low number of recovered VSGs might be due to the quality of the initial data, which affects the quality of the genome assembly, leading to short contigs. Due to the length thresholds applied to the sequence similarity search (≥ 150 amino acids), only contigs longer than 450 nucleotides were considered. The average full length of the reference VSGs is 404 amino acids, whilst the average

length of the VSG recovered per strain ranges from 190 to 340 amino acids, corresponding to 47 % to 86 % of the VSG average full length.

To assess whether the VSGs were sufficient to produce reliable VAPs through the same methodology as *T. congolense* VSG profiling, a sampling exercise was performed on the reference full repertoire. The original VSG repertoire of TvY486 has 723 VSG-like sequences. These VSGs were divided into 96 phylotypes based on their genetic distances. To achieve this, the TvY486 VSG phylogenies were inspected and each clade was assigned a phylotype number. Consecutively smaller percentages of the repertoire were sampled and VAPs produced to observe whether all 96 phylotypes in the reference repertoire were recovered. Random sampling revealed that to obtain all phylotypes, the VSG repertoire had to be complete. This is because 69 of the 96 phylotypes (72 %) contain less than 10 genes. Therefore, the chance of recovering any particular phylotype is low because they are infrequent. This, coupled with the large number of phylotypes relative to *T. congolense*, means that 100 % of genome coverage is necessary to capture all VSG sequences. This contrasts with the situation in *T. congolense*, where only 15 % of the VSG repertoire was sufficient to provide an accurate profile.

Sampling a small part of the VSG repertoire only gives a robust estimate of sub-familial proportions (**Table 13**). When looking at the percentage of Fam23-26 compared to the reference, no significant difference between the reference full repertoire and the field isolates repertoires was found (**Figure 46**), suggesting that the abundance of each family is both highly conserved and different enough to be detected even at low sequencing depth.

This means that the VAPs cannot be accurately estimated from the data simply using the same sequence similarity-based methodology as *T. congolense*.

Table 13 Number of VSG recovered from each strain by sequence similarity search and their proportion compared to the reference full repertoire.

Sample ID	Fam23	%	Fam24	%	Fam25	%	Fam26	%	TOTAL	%	Average Length
TvY486	363	50.35%	161	22.33%	133	18.45%	64	8.88%	721	100%	404
IL11	112	45.34%	46	18.62%	63	25.51%	26	10.53%	247	34.26%	349
IL1392	132	43.56%	65	21.45%	75	24.75%	31	10.23%	303	42.02%	305
IL2005	28	52.83%	14	26.42%	6	11.32%	5	9.43%	53	7.35%	207
IL2323	73	54.89%	41	30.83%	11	8.27%	8	6.02%	133	18.45%	305
IL2714	83	56.46%	42	28.57%	15	10.20%	7	4.76%	147	20.39%	305
IL306	73	40.11%	38	20.88%	52	28.57%	19	10.44%	182	25.24%	281
IL3171	26	65.00%	6	15.00%	6	15.00%	2	5.00%	40	5.55%	187
IL319	77	42.78%	35	19.44%	52	28.89%	16	8.89%	180	24.97%	221
IL338	81	45.51%	41	23.03%	39	21.91%	17	9.55%	178	24.69%	220
IL340	72	49.32%	31	21.23%	34	23.29%	9	6.16%	146	20.25%	207
IL3638	23	44.23%	15	28.85%	10	19.23%	4	7.69%	52	7.21%	309
IL3651	44	64.71%	15	22.06%	9	13.24%	0	0.00%	68	9.43%	332
IL3658	45	51.14%	16	18.18%	16	18.18%	11	12.50%	88	12.21%	339
IL462	87	47.54%	38	20.77%	37	20.22%	21	11.48%	183	25.38%	275
IL465	69	44.81%	30	19.48%	41	26.62%	14	9.09%	154	21.36%	275
IL493	70	46.05%	29	19.08%	42	27.63%	11	7.24%	152	21.08%	220
IL596	78	46.71%	31	18.56%	44	26.35%	14	8.38%	167	23.16%	219
IL684	96	45.71%	46	21.90%	51	24.29%	17	8.10%	210	29.13%	226
ILV-21	22	52.38%	14	33.33%	3	7.14%	3	7.14%	42	5.83%	190

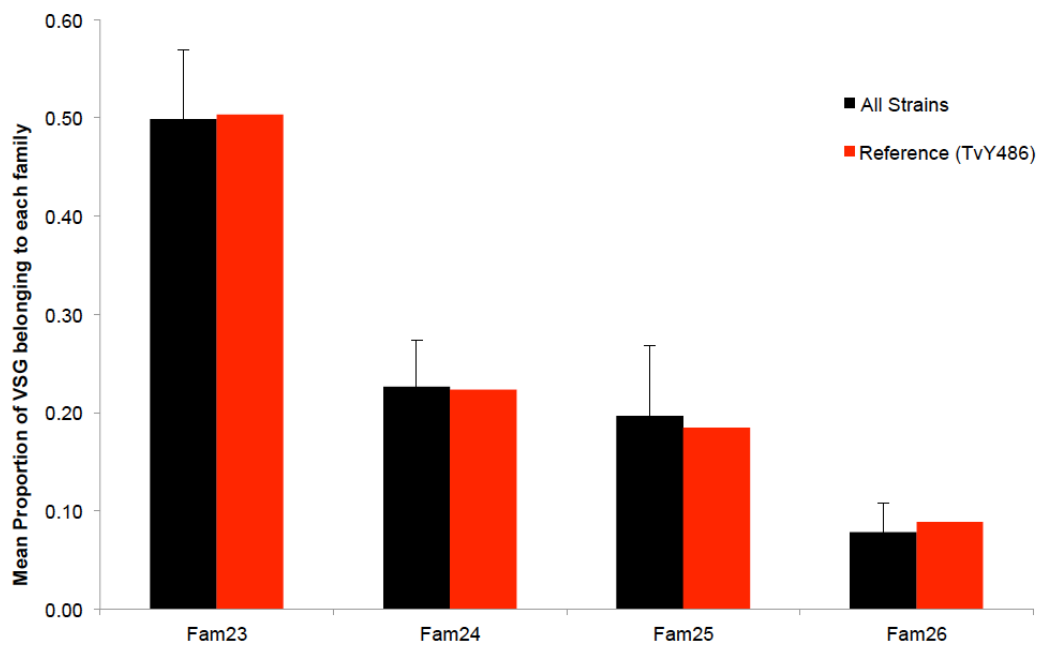


Figure 46 Proportion of VSG belonging to families 23-26 from all field strains compared to the proportions of the reference full VSG repertoire given as mean \pm σ . Bars are colour coded according to key.

Due to the high number of phylotypes and their low frequency, the genomes are never complete enough to produce accurate VAPs based on the phylotype frequencies, as done in chapter 2 for *T. congolense*. Low abundance phylotypes would be represented by frequencies too low to measure variation. Therefore, I have adopted an approach based on the presence of diagnostic COGs. Whilst in *T. congolense*, the VAP relies on shared patterns within the VSG sequences; in *T. vivax* it identifies VSG sequences that are characteristic of particular locations.

To pursue this method, family phylogenies were estimated using all recovered VSGs and the reference full repertoire to identify new phylotypes, absent in the reference genome. **Figure 47** shows ML phylogenies for Fam23-26. The reference nodes are highlighted in red, showing that sequences from the field isolates cluster amongst the reference sequences within each phylotype. This indicates that the cladistic structure of all four families is conserved across the strain cohort. No novel phylotypes were identified. Furthermore, 95 out of the 96 phylotypes identified in the reference genome were present in at least two other isolates. These results indicate that the majority of phylotypes are widespread. In contrast to *T. congolense*, in *T.*

vivax the data cannot confirm universality because of the low numbers of VSGs recovered and the smaller geographic distribution of the sample cohort.

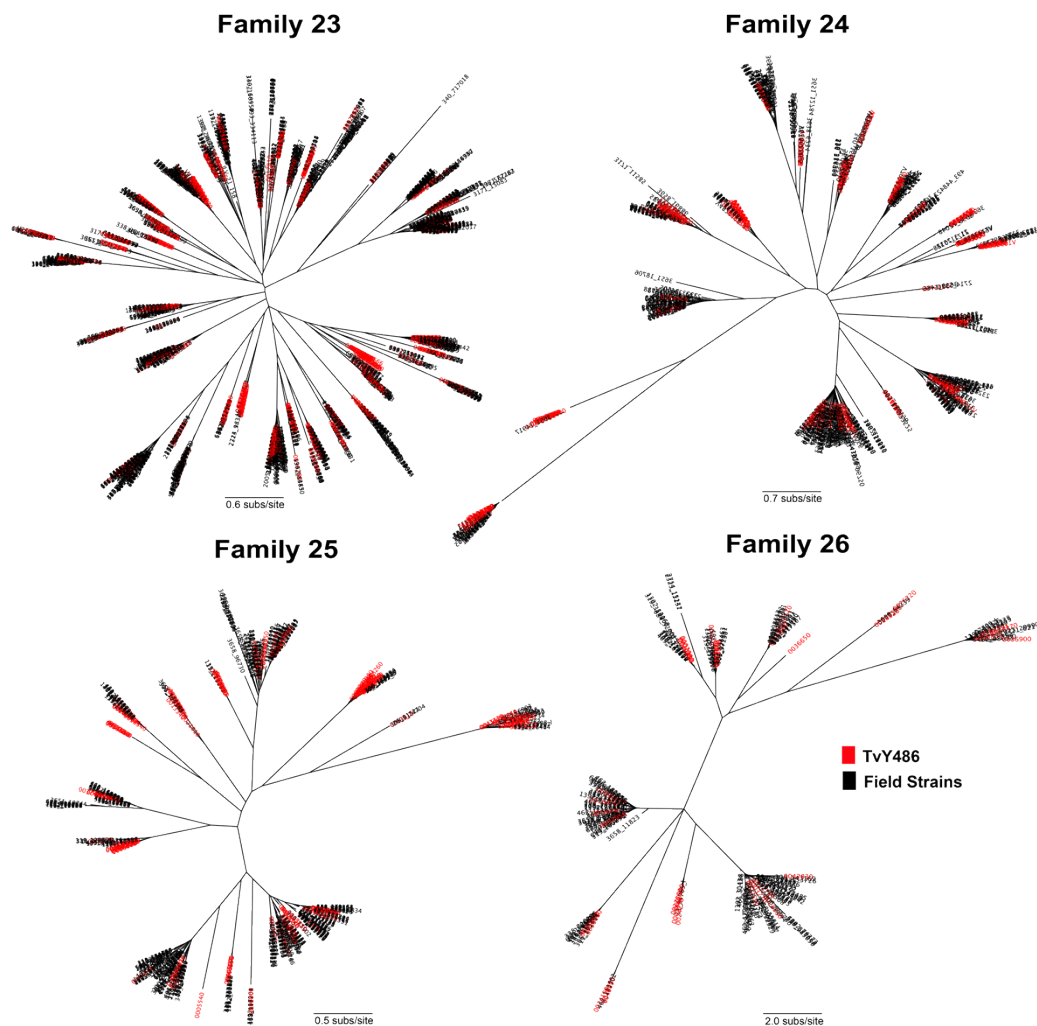


Figure 47 Phylogenies of the VSG families. Reference strain sequences highlighted in red, field strain sequences are shown in black. The phylogenies show that the reference cladistic structure is maintained in the remaining isolates. Unrooted trees were estimated with maximum likelihood method and the WAG+Γ model using RaxML (Stamatakis 2014).

5.3.3 Intra-phylotype variation

To identify VSG orthologs, phylogenies were built for each phylotype with 4 or more sequences (all except phylotypes 29, 30, 69 and 83); COGs were identified based on their phylogenetic position and their sequence identity (>98 %). The longest genes of each of the COGs were compiled in a database. Genes belonging to

phylotypes with less than 4 sequences were subsequently added to the database as individual entries.

The purpose of the COG database is to screen VSG contigs or VSG reads and identify shared sequences. Whilst this approach is efficient at associating VSG sequences that are present in the database, it fails to identify novel VSGs that were not detected in the sample cohort. Therefore, until the *T. vivax* genome diversity has been extensively sampled, new genomes may have to be searched for novel, strain-specific VSG to update the database.

The COG database was used to screen VSG using two different approaches: first, by searching contigs based in sequence similarity with 95 % and 98 % identity thresholds and, second, by read mapping under two sensitivity thresholds (**Table 14**). In both approaches, the number of COGs recovered depends on the mapping or identity threshold. The most stringent approach is contig search at 98 % identity (463 ± 269 COGs recovered), whilst the most flexible is read mapping under “very sensitive” settings (710 ± 244 COGs recovered) (**Table 14**). This difference results from size of compared sequences and data availability. First, the shorter size of reads compared to contigs means that a smaller matching sequence is required for a positive identification. Second, read mapping considers more data than contigs, because of data lost during the genome assembly process.

Table 14 Distribution of COGs recovered by the two approaches under different sensitivity thresholds. Contigs refers to results from sequence similarity searches with 98 % and 95 % sequence identity thresholds. Reads refers to read mapping with Bowtie2 using the sensitive or very sensitive options.

	Contigs (98% ID)	Contigs (95% ID)	Reads (Sensitive)	Reads (Very Sensitive)
Total (Mean $\pm\sigma$)	463 \pm 269	558 \pm 242	708 \pm 244	710 \pm 244
All Locations	48	103	246	250
All but The Gambia	52	145	240	226
All but Uganda	71	83	98	102
Nigeria only	416	110	138	173
Ivory Coast only	69	29	0	0
The Gambia only	8	5	0	0
Uganda only	0	53	0	0
Other Combinations	299	531	351	365

In the contig similarity search with 98 % identity threshold, 48 COGs were found in all isolates from all locations (**Table 14**). These correspond to VSG that are conserved across the species. A total of 52 COGs were found in isolates from all

locations except from The Gambia. The sample cohort only has one isolate from The Gambia (IL3171), which has low genome coverage (55 %, shown in **Table 12**). Therefore, it is possible that some or all of these 52 COGs are present in Gambian isolates, but have been missed in this dataset. Seventy-one of COGs were found in isolates from all locations except in any of the four Ugandan isolates. Although only 6 to 20 % of the VSG repertoire of these isolates has been analysed, it is possible that they also represent missing data. However, these 71 COGs are absent from all four isolates, not just one as in the previous example from The Gambia. This approach also identified 416 COGs apparently specific to Nigeria, 69 to Ivory Coast, and 8 to The Gambia. The high number of Nigeria-specific COGs may reflect the high number of Nigerian samples included in the dataset (N = 11) and the fact that their sequencing output was generally better [77 to 80 % of genome coverage (**Table 12**) and up to 42 % of the VSG repertoire recovered (**Table 13**)]. As there are more sequences from Nigerian samples than the other locations, it is possible that there is a bias in the sequence search. Furthermore, this is supported by the fact that when the sequence similarity search identity threshold is reduced to 95 %, the number of Nigeria-specific COGs reduces by 74 % to 110 (**Table 14**).

In the read mapping approach, more COGs are recovered from the genomes in both sensitive and very sensitive settings than with the contig sequence similarity search (463 ± 269 COGs recovered with contig search at 98 % identity; 710 ± 244 COGs recovered with read mapping under very sensitive settings) (**Table 14**). As the difference in the number of VSG recovered between sensitive and very sensitive settings is not statistically significant (p-value = 0.43) and the latter increases the run time, the sensitive option is preferred. With read mapping (sensitive), 246 COGs were found in isolates from all countries. No COGs were found to be specific for Ivory Coast, The Gambia or Uganda, whilst 138 COGs were found exclusively in Nigerian isolates.

COG analysis is useful to identify diagnostic markers for geography and disease phenotype, but also to discriminate amongst antigens of different isolates. With the available data, it is possible to select region-specific genes. However, their nature depends on the approach taken. Choosing the most stringent method (contig mapping, 98 % ID), results in a large cohort of country-specific COGs (N(Nigeria) = 416, N(Ivory Coast) = 69, N(The Gambia) = 8). Together they represent 45.6 % of the COG database. Taking into account multiple-country combinations, the data suggest that only 2.3 % of the VSG pool is common to all African *T. vivax*. Even if

the Gambian isolate is disregarded due to the reasons discussed above, this number only increases to 9.2 %. However, if a more relaxed approach is taken (read mapping, sensitive settings), the percentage of widespread COGs increases to 44.9 % (N = 486), representing the majority of phylotypes (84/96). The phylotypes not represented are mostly of single-copy genes (7/12), and thus their absence could relate to incomplete sampling as noted in section 5.3.2.

Together, these observations suggest that at most 84 phylotypes are universal to the *T. vivax* strains used in this study. Furthermore, when comparing the COG database to the PacBio genome of TvY486 by read mapping, all COGs can be found, including those that were not described in the original TvY486 genome produced by Sanger sequencing. Therefore, the read mapping approach using the new genome as reference reduces number of country-specific genes to only 138 (all from Nigeria – the original location of TvY486). Nevertheless, using either approach, a large part of the VSG repertoire (27.6 % to 49.1 %) is conserved across multiple locations at a very high sequence identity (> 98 %), suggesting that the *T. vivax* VSG repertoire is stable, remaining conserved across isolates in the same manner as the core genes. Better genome coverage and further sampling are likely to reinforce this argument by reducing the effect of genome fragmentation and incomplete geographical coverage.

5.3.4 Genetic relationship between strains

To understand the genetic relationship between strains, a minimum evolution phylogeny was estimated based on whole-genome SNPs (**Figure 48A**). All isolates grouped according to geographic region of collection. Cluster analysis of the recovered COGs as a binary matrix also showed a similar relationship between strains with the exception of isolate Tv3651 from Ivory Coast (**Figure 48B**). In the cladogram estimated from the COG matrix, Tv3651 clusters with the Ugandan strains. The direct correlation between the COG matrix and the whole genome variation also support the hypothesis that the *T. vivax* VSG repertoire diverges similarly to the core genome.

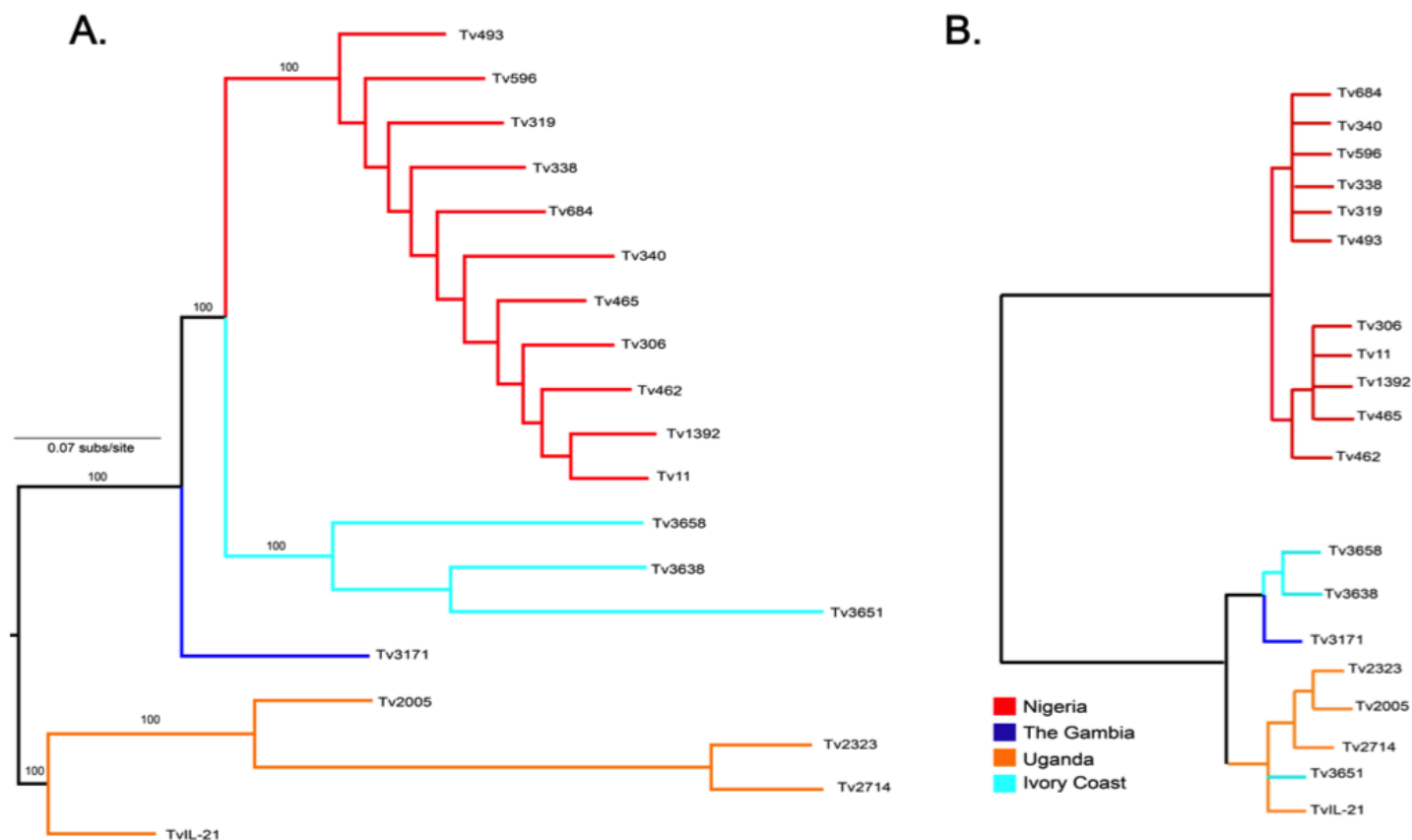


Figure 48 The relationship between *T. vivax* SNPs and COGs. A. Maximum likelihood phylogeny of *T. vivax* whole genome SNPs. Tree was estimated with RAxML, mid-rooted. 0.07 nucleotide substitutions per site. B. Dendrogram of *T. vivax* strains based on presence and absence of 1081 COGs. Colour-coded by country of collection according to key.

5.3.5 The genomic VSG repertoire of *T. vivax* Lins

To show how the VAP can be applied to characterise clinical isolates, the VSG repertoire of the Brazilian strain *T. vivax* Lins was analysed. For such study, genomic reads were screened for a collection of VSG orthologues retrieved from the population genomics analysis of African *T. vivax* presented in the previous section. A heatmap and a cladogram were produced based on presence and absence of these orthologues (**Figure 49**). TvLins clusters within 3 Ugandan samples, being each other's closest relatives.

To corroborate whether the genetic affinity of TvLins to the Ugandan isolates was a feature of the whole genome or just the VSG repertoire, a whole genome SNP phylogeny was estimated. This showed a defined geographical signature across Africa and verifies that the Ugandan strains are TvLins' closest relatives in this sample cohort (**Figure 50**). This situation contrasts with the previously published Venezuelan isolate TvLIEM (Greif et al. 2013), which is closer to the Gambian isolate Tv3171.

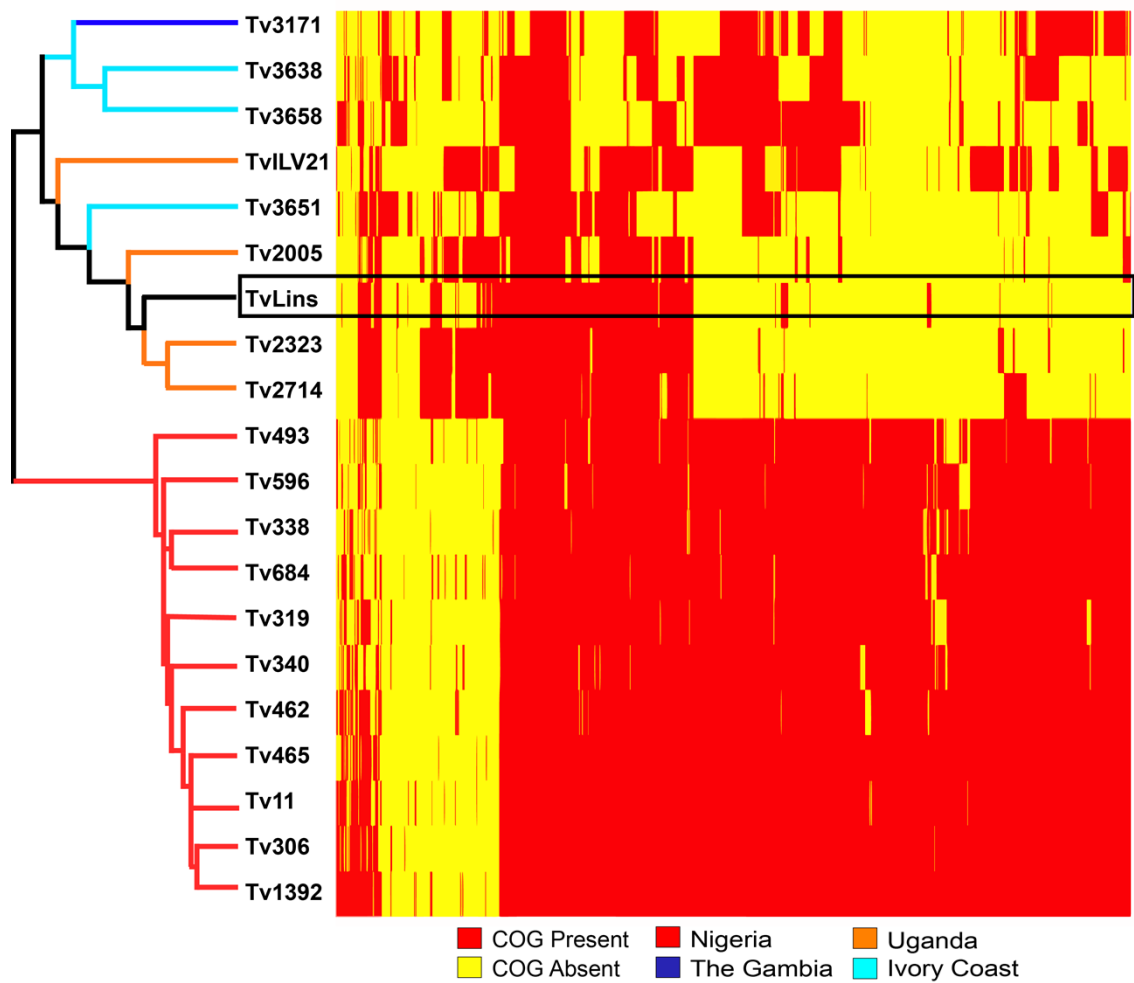


Figure 49 Heatmap and cladogram showing presence and absence of 1081 VSG orthologues across our sample dataset. In the heatmap, presence of COGs is indicated in red, and absence indicated in yellow. In the dendrogram, strains are coloured by their geographic location. Positioning of TvLins is indicated by the black rectangle.

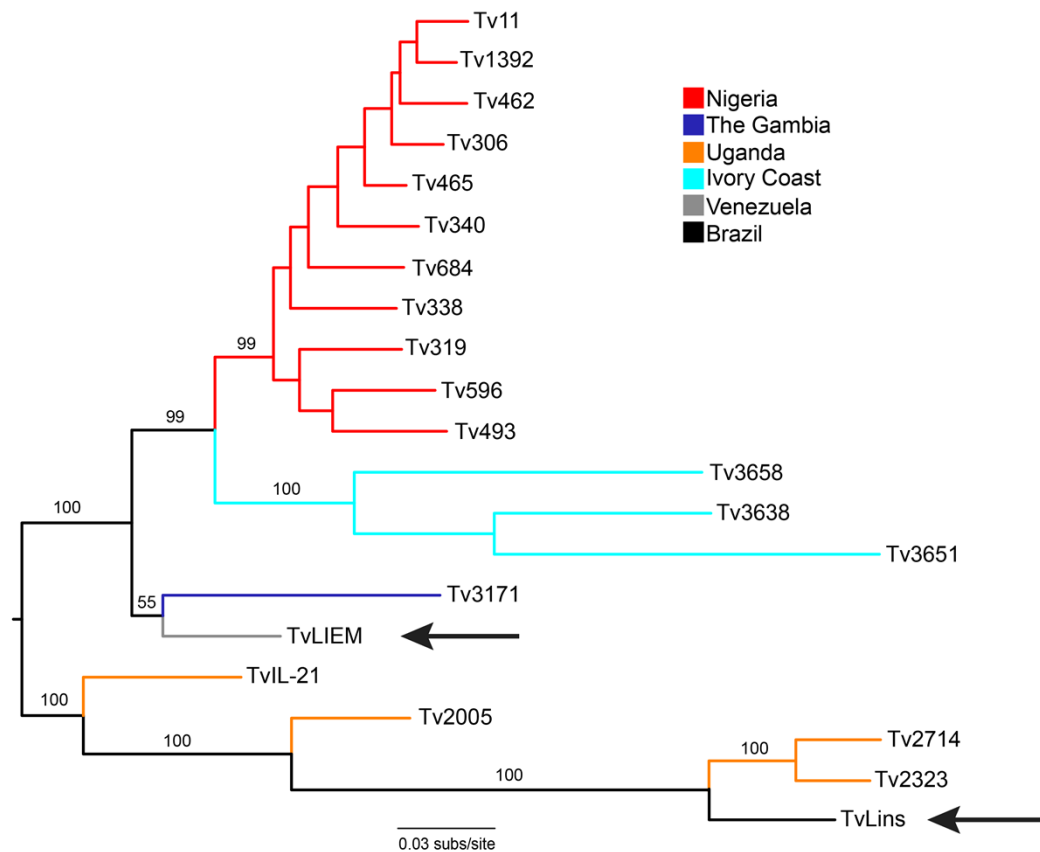


Figure 50 Maximum likelihood phylogeny of *T. vivax* strains based on whole genome SNPs estimated with GATK (N = 1,011,378). SNPs called from the transcriptome of the Venezuelan isolate TvLIEM were added posteriorly to the analysis (N = 7,848 SNPs) to include an independent representative of South American *T. vivax* (Greif et al. 2013). Tree is mid-rooted. Strains are colour coded according to geographical location. Arrows indicate samples from South America (TvLIEM and TvLins).

As noted earlier, the COG database cannot find novel VSGs. Therefore, to check whether TvLins contained specific VSG, VSG contigs were retrieved by sequence similarity search, using a database of TvY486 VSGs. A total of 409 VSGs were recovered. All 4 VSG families (Fam23-Fam26) were represented in similar proportions as to both the reference and the historical isolates (**Figure 51**). The recovered VSGs were aligned with all the VSGs from TvY486 and the four Ugandan strains (**Figure 52**). The Ugandan samples were included because they are closer to TvLins than TvY486. Therefore, they can be used to identify VSGs absent in the reference but not specific to this Brazilian strain. In all families, all VSG nodes are closely related to the Ugandan samples and/or TvY486 and no novel VSGs were

detected. Clusters of TvLins-specific sequences were never observed, suggesting that the COG database is sufficient to describe the VSG repertoire of this isolate.

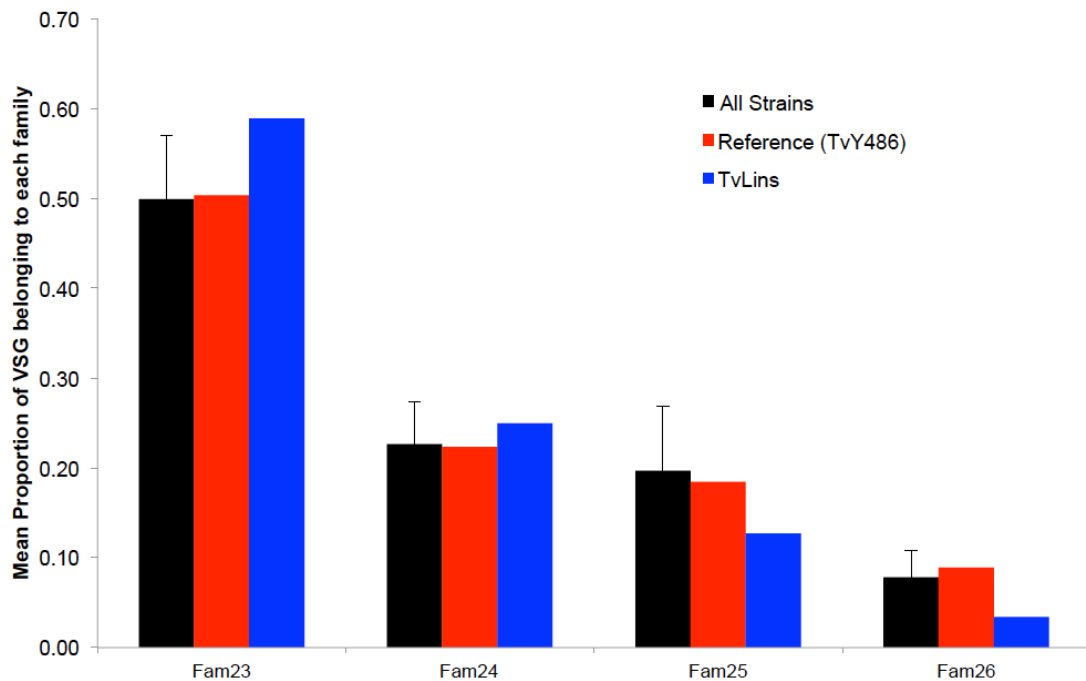


Figure 51 Relative Frequency of each VSG family in our strain dataset, TvY486, and TvLins. Values for all strains are mean \pm σ . Bars are colour-coded according to key.

In summary, the VSG repertoire in TvLins reflects the repertoire of the African *T. vivax* samples. However, unlike previous reports in literature and the Venezuelan TvLIEM (Cortez et al. 2006; Greif et al. 2013), it suggests that this strain is closer to Ugandan strains (East Africa) than to West African strains, including the reference TvY486.

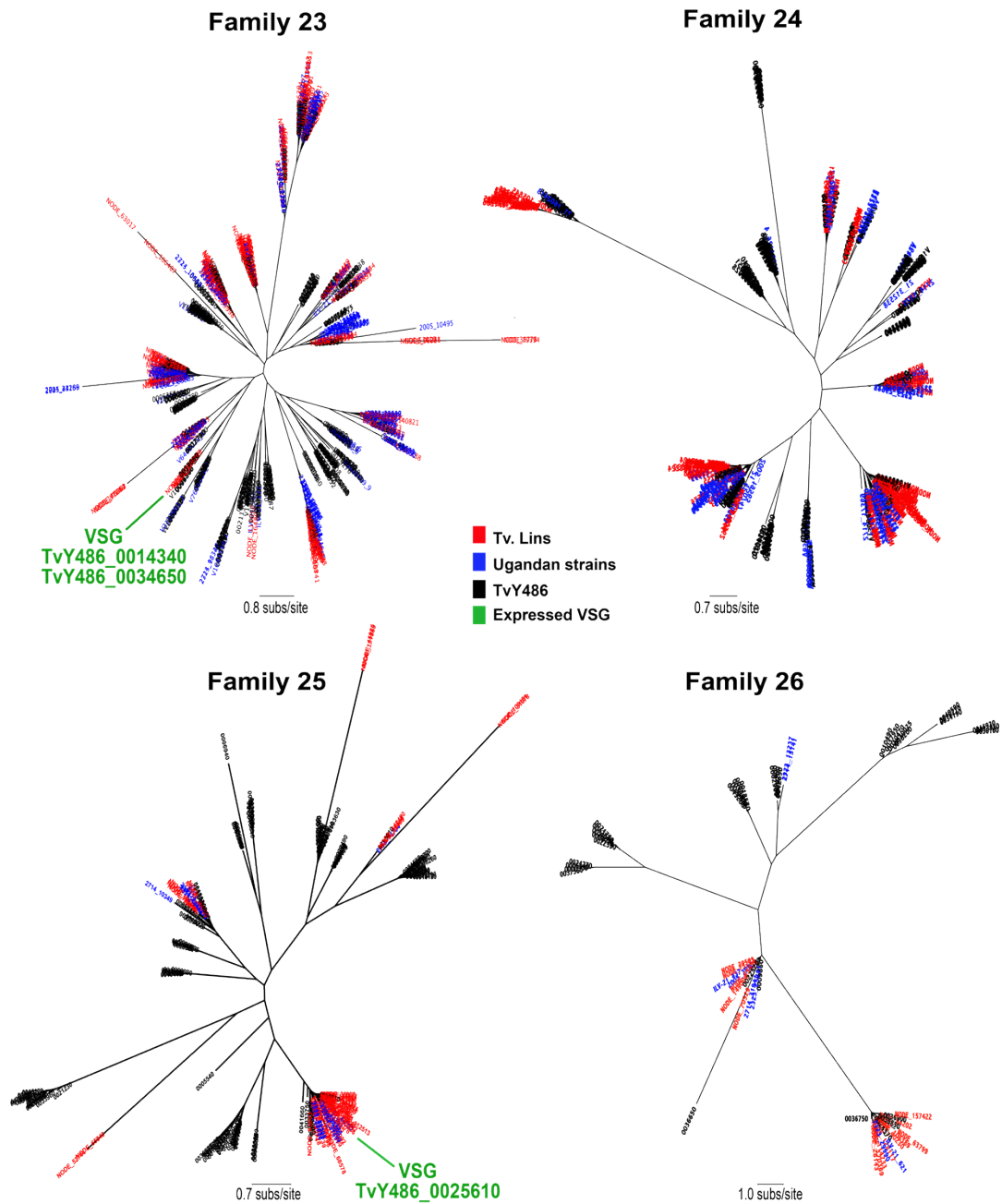


Figure 52 Maximum likelihood phylogeny of amino acid sequences of VSG families 23-26 using WAG+F model of amino acid substitution. Nodes in black correspond to sequences of the reference Y486, blue of Ugandan samples, while red shows sequences from *T. vivax* Lins. The green labels identify the most abundant VSG transcripts seen in the transcriptome replicates S3 and S4.

5.3.6 The expressed VSGs in *T. vivax* Lins

The VSG database can also be applied to transcriptomic data to compare VSG expression between isolates or different time points. Two transcriptomes were produced from replicate experimental infections in goats, named S3 and S4. Reads were mapped to the reference TvY486 with Bowtie2 (Langmead & Salzberg 2012). Subsequently, cufflinks (Trapnell et al. 2012) was used to estimate transcript abundances. The transcriptomic profile showed that the most abundant transcripts were VSGs, ribosomal proteins and cytoskeleton-associated proteins. A total of 94 and 89 VSG transcripts were recovered from S3 and S4, respectively. The FPKM values ranged from 0.1 to 25073.

The transcriptome results were compared to the bloodstream transcriptome of IL1392, previously published by Jackson et al. (2015) (**Table 15**). The IL1392 transcriptome was produced from parasites maintained *in vivo* by mouse passage. The three transcriptomes were produced from RNA extracted from parasites purified from blood by differential centrifugation as described in Jackson et al. (2015). Although the IL1392 transcriptome has more read pairs, the percentage of transcripts mapped to the reference genome are comparable [89.77 %, 90.35 %, and 84.24 % for IL1392, S3 and S4, respectively (**Table 15**)].

Table 15 Summary of VSG transcripts in S3 and S4 compared to the IL1392 transcriptome previously published by Jackson et al. (2015).

	IL1392	S3	S4
Number of Reads	14175293	6722702	7452591
% Map to the TvY486 genome	89.77%	90.35%	89.24%
Number of VSG transcripts	65	94	89
FPKM range	[0.08-16462.35]	[0.09- 25073.7]	[0.11-954.48]

The average FPKM for the 1000 most abundant transcripts of S3 and S4 is lower than for IL1392 (Mean: S3 = 247, S4 = 282, IL1392 = 547; Median: S3 = 145, S4 = 165, IL1392 = 224). The differences in transcript abundance between both S3 and S4 and IL1392 are statistically different (Independent t-test, $p < 0.001$), but the differences between S3 and S4 are not (Independent t-test, $p = 0.65$). To further understand the differences in transcript abundance between samples, FPKM

correlations were calculated. In the first panel, **Figure 53** shows the positive correlation of FPKM values between S3 and S4 ($R^2 = 0.99$), indicating that the samples are good replicates. The second and third panels show the correlation between the IL1392 transcript abundances and the ratio of the abundances of TvLins and IL1392 ($R^2(\text{S3}) = 0.88$, $R^2(\text{S4}) = 0.92$). The data was transformed to show the exponential correlation between S3/S4 and IL1392. These results indicate that transcripts of higher abundance in IL1392 have a weaker correlation with both S3 and S4. Such differences are likely to be a reflection of the different sequencing effort because read coverage affects high abundance transcripts more than low abundance transcripts. This is corroborated by the higher number of reads in IL1392 compared to S3 and S4 as observed in **Table 15**.

Regarding the VSG repertoire, despite the higher number of reads, more VSG transcripts were recovered from S3 and S4 than from IL1392 (94 and 89) (**Table 15**). However, only in S3 was a super abundant VSG found (FPKM = 25,073.7). It is a co-orthologue of the reference gene TvY486_0025610, and belongs to Fam25. This shows for the first time that Fam25 encodes functional VSGs. In sample S4, a super abundant VSG could not be identified, but two VSGs from Fam23 were the most abundant (co-orthologous to TvY486_0014340 and TvY486_0034650; FPKM values of 932.2 and 732.3, respectively). The phylogenetic positions of the three most abundant VSGs are indicated in green in **Figure 52**. The six VSGs with highest FPKM for each sample were compared to the IL1392 *T. vivax* transcriptome and are shown in **Figure 54**. This highlights the difference in magnitude between the super-abundant VSG and the others, and indicates that VSGs can be characterised using the VAP. Here, it is shown that within the 6 most abundant VSGs of each sample, there are members of Fam23, Fam24 and Fam25.

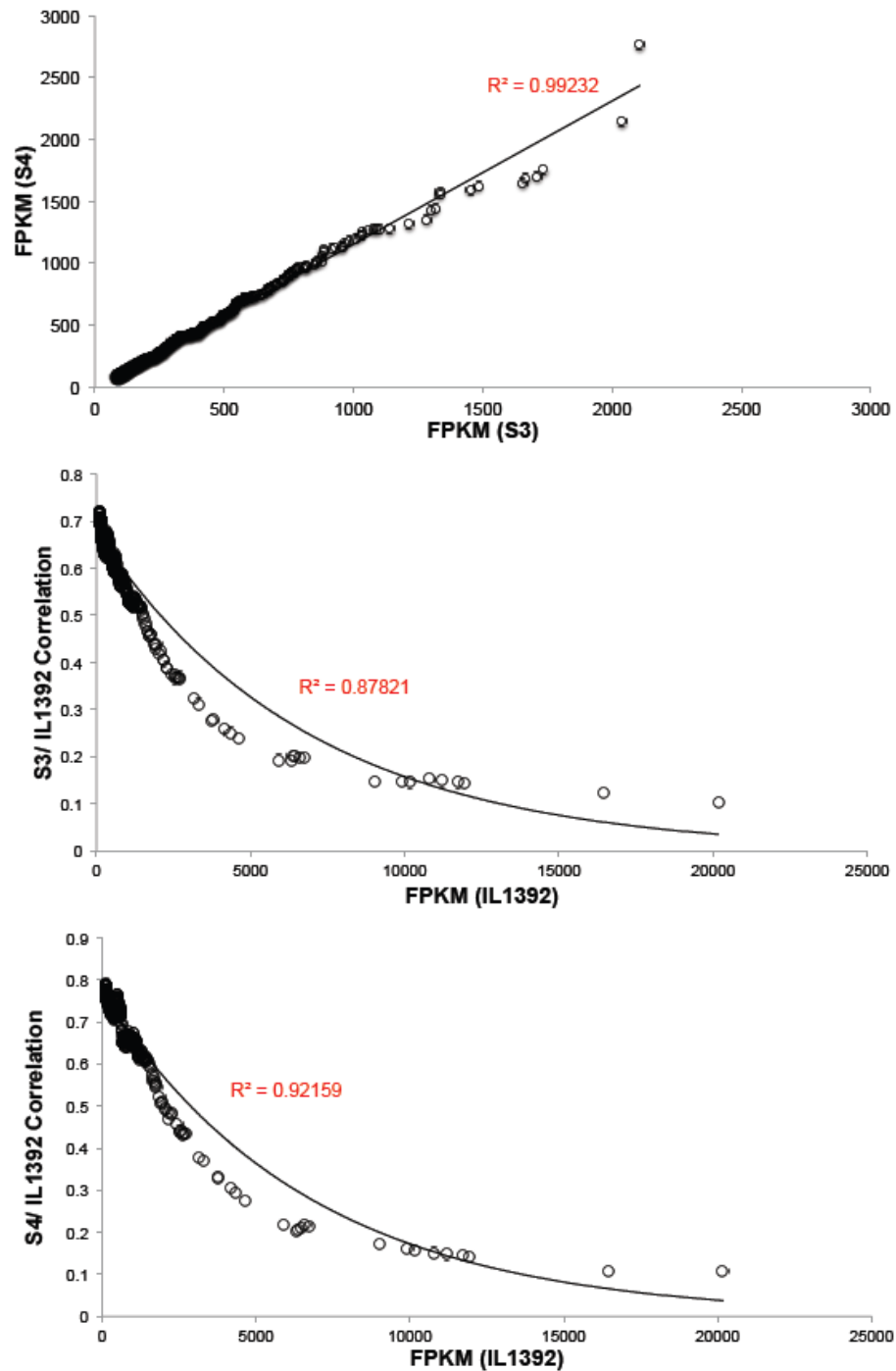


Figure 53 Correlation of transcript abundances based on the 1000 most abundant transcripts between S3, S4, and IL1392. The superabundant VSGs of S3 and IL1392 were removed. The first graph shows the linear correlation between S3 and S4 transcript abundance values, expressed as fragments per kilobase of transcript per million mapped reads (FPKM). The second and third graphs show the exponential correlation between the ratio of FPKM values of S3/S4 and IL1392 respectively. R^2 values are shown in red ($p < 0.05$).

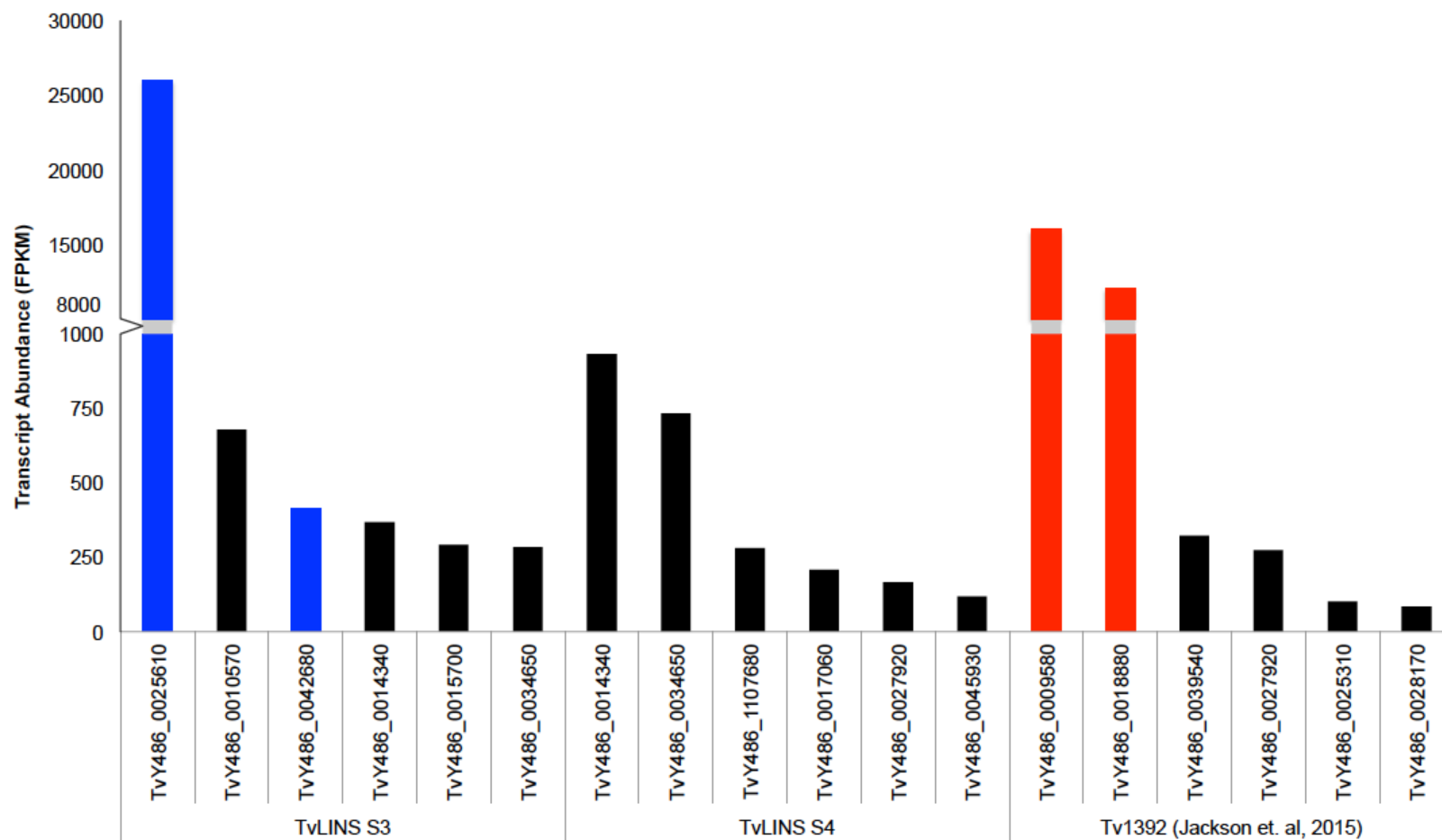


Figure 54 FPKM values of top 6 transcripts and correspondent VSG families for S3. Transcripts belonging to Fam23 are shaded in black, Fam24 in red and Fam25 in blue.

5.4 Discussion

This chapter presented the first study of *T. vivax* VSG diversity on a population scale. I developed a VAP methodology suitable for the peculiarities of the *T. vivax* VSG repertoire, based on the presence or absence of diagnostic VSG orthologs. As in chapter 5, 13 of the 20 isolates used here were passaged into naïve rodents previous to DNA extraction, which may have reduced the diversity of the study. This is particularly pertinent in *T. vivax*, as the majority of isolates is thought not to grow in rodents (Morrison et al. 2016). However, this could not be avoided, not only because of the barely undetectable parasitaemia characteristic of *T. vivax* infections in endemic areas, but also of the challenges of finding cryopreserved *T. vivax* blood stabilates. In the second part of this chapter, I have applied the VAP to the bloodstream stage transcriptome of TvLins to show that the VAP can successfully describe VSG expression.

In chapter 2, I developed the VAP based on conserved protein structural motifs that successfully differentiate the 15 phylotypes that comprise the *T. congolense* VSG repertoire. Here, I had to adopt a different methodological approach because the *T. vivax* VSG repertoire is structurally very distinct from *T. congolense*. First, the *T. vivax* VSG repertoire is comprised of 96 low-abundance phylotypes, which precludes accurate comparisons of phylotype relative frequencies. Since all phylotypes are a small proportion of total VSGs, this diversity also means that a genome sequence must be almost complete for all phylotypes to be detected. Recovering the full VSG repertoire is challenging when sequencing field isolates due to the low parasitaemia of natural infections, but even in experimental infections, the sequencing depth and coverage would have to be very high to recover every phylotype frequency and be sure that unrecorded phylotypes were genuinely absent. Second, whilst in *T. congolense* the 15 VSG phylotypes are universal and exhaustive, in *T. vivax* I could not establish universality. Whilst the evidence supports that most VSGs are conserved in the population, I found VSGs specific to particular countries. Whether they are all real or result from incomplete sampling remains unresolved. With these data, I presented a novel VAP approach that, rather than measuring conservation across the repertoire as is done for *T. congolense*, detects and relates diagnostic VSGs. At present, these VSGs are associated with the isolate geographical location. In future, I envisage these

markers to relate to other aspects of epidemiology, particularly host use, disease outcome, and virulence.

5.4.1 Antigenic diversity in *T. vivax*

In this chapter, I have shown that the *T. vivax* VSG repertoire has 96 lineages, distributed by four structurally distinct families that are conserved in the population in stable proportions. I further showed that, whilst there are examples of geographically restricted VSGs, these lineages are conserved in the twenty genomes from Africa and South America analysed here. This contrasts with both *T. brucei* and *T. congolense*. The VSG repertoire of *T. brucei* is very dynamic, resulting in the development of antigenically distinct ‘mosaic’ VSGs even during the course of a single infection (Hall et al. 2013). In *T. congolense*, the phylotype proportions are stable across isolates, but there are significant fluctuations that allow differentiation among isolates. For example, I showed in chapter 2 that the Gambian *T. congolense* isolates had a discrete signature of larger phylotypes 5 and 6 (page 64). Moreover, in *T. congolense* the VSG repertoire stability is reflected in the phylotype proportions alone, whilst in *T. vivax*, the repertoire conservation can be detected to the level of individual genes or orthologue groups.

The conservation in the *T. vivax* VSG repertoire may reflect clonality. In the whole genome SNP analysis, I show little overlap between populations of different countries and little evidence for genetic exchange. This is consistent with the evidence for asexual reproduction in this species suggested by Duffy et al. (2009). In their paper, 531 samples were collected from The Gambia between March 2006 and January 2007 from donkeys, horses and cattle and analysed by multi-locus genotyping. The *T. vivax* genome encodes meiosis-associated genes, but its population structure is consistent with clonal expansion, showing reduced genetic diversity, clonal expansion of particular genotypes, and significant levels of linkage disequilibrium (Duffy et al. 2009). Together, these data suggest the parasite has lost the ability to mate after separating from the *T. brucei*/*T. congolense* ancestor (Duffy et al. 2009).

Clonal expansion may also explain the direct relationship between VAP, population structure and geography. The sample cohort retains a strong geographic signal at the genomic level that is reflected by the VSG repertoire (**Figure 48**). In contrast, in

T. congolense, the correlation between the VAP and the population structure is weak. In *T. brucei*, there is a marked genetic difference between West African *T. brucei gambiense* and East African *T. brucei rhodesiense* (Gibson 2001). Even within *T. brucei rhodesiense* and *T. brucei gambiense* populations, geography is often a main source of genetic variation (Duffy et al. 2013; Echodu et al. 2015). For example, long phylogenetic branches separate isolates from Ivory Coast, Guinea and Cameroon, even when more than one focus exists in the country and collection was spread over 50 years (Weir et al. 2016). However, mating frequency was still shown to influence population structure. This is evidenced by the heterogeneity of *T. brucei rhodesiense* populations in Malawi, where mating was recurrent, as opposed to the lower genetic diversity of Uganda, where evidence for genetic exchange is scarce (Duffy et al. 2013).

The high number of conserved phylotypes is unprecedented in organisms employing antigenic variation. In *T. brucei*, VSGs have only been divided into a and b-type, although the N-termini can be split into 3 groups (A-C) and the C-termini into six groups (1-6) (Carrington et al. 1991; Marcello & Barry 2007a). Similarly, in *T. congolense*, VSGs can be divided into 15 lineages (Jackson et al. 2012). In *Plasmodium falciparum*, antigenic variation occurs mainly through the *var* genes, a family of variant surface antigens encoding PfEMP1 (Baruch et al. 1995). PfEMP1 is involved in virulence, including but not exclusively through the mediation of parasite adhesion to host endothelial receptors (Lavstsen et al. 2003). Although vast, the *var* repertoire of *P. falciparum* is divided into only three major groups (group A, B and C) and two intermediate groups B/A and B/C representing transitions (Lavstsen et al. 2003; Kraemer & Smith 2003). The three major groups are comparable to the VSG families Fam23-26 of *T. vivax*, which raises an important distinction in gene repertoire: whilst *var* groups A-C can interchange genes, giving rise to intermediate groups B/A and B/C, the four VSG families in *T. vivax* are too structurally distinct for recurrent gene conversion. Thus, *T. vivax* is not only conserving structurally distinct families, but also conserving numerous phylotypes within these families and a defined cladistic structure within those phylotypes.

Similarly to *T. congolense*, phylotype conservation may have functional reasons. For example, different phylotypes, or even individual VSGs, may determine clinical outcome. As the VAP also delivers a method to analyse transcriptomes, it is now possible to examine VSG expression in isolates and conditions. This is evidenced by the analysis of the TvLins transcriptome. The VSG repertoire of TvLins is not

atypical and all four families are represented in similar proportions as the African strain collection. However, the profiling of the two transcriptome replicates, S3 and S4, does reveal novel results, since a member of Fam25 has for the first time been reported as the superabundant VSG. Fam25 and Fam26 were first described in the African trypanosome phylome study as VSG-like families not belonging to either a-type or b-type VSG and with no co-orthologues in either *T. brucei* or *T. congolense* genomes, which introduced doubt about their function as variant antigens (Jackson et al. 2013). The TvLins transcriptome results support the hypothesis that Fam23, Fam24 and, at least, Fam25 encode functional variant antigens.

Understanding the diversity of the *T. vivax* VSG repertoire may help explaining the phenotypic variability of *T. vivax* infections. For example, in Brazil, parasite genotypes circulating in five farms from two municipalities in Pantanal were found to cause mild, chronic disease, causing symptoms exclusively as a secondary pathogen (Paiva et al. 2000). However, in Pernambuco, *T. vivax* has led to symptomatic disease, characterised by reduced milk production, premature offspring and miscarriages (de Souza Pimentel et al. 2012). In the first *T. vivax* outbreak in the state of São Paulo, caused by the TvLins isolate, disease was very acute, triggering a wide range of haematological and neurological symptoms (Cadioli et al. 2012). Similarly, in Paraíba, a *T. vivax* outbreak has caused the loss of 71 % of a sheep herd, even though cattle and buffalos of the same farm had only mild or asymptomatic disease (Galiza et al. 2011). These studies suggest that the clinical outcome of *T. vivax* infections is diverse and unpredictable in Brazil, something that has long been observed in Africa (Hornby 1921; Hudson 1944; Losos & Ikede 1972). It is possible that the different disease outcomes reflect different host-parasite interactions. Thus, as the main surface proteins interacting with the host's immune system, the VSGs may play a role in infection phenotype. The VAP provides a method to start investigating such associations.

If particular VSGs can be linked to disease outcome, geographical location, or host susceptibility, the VAP becomes an epidemiological and diagnostic tool. Diagnosis of HAT and surra based on variant antigens already exists. For instance, *T. brucei rhodesiense* infections can be diagnosed by a SRA-specific PCR (Radwanska et al. 2002). For *T. brucei gambiense*, there are two rapid diagnostic tests targeting a combination of VSGs (HAT Sero-K-SeT and HAT Sero-Strip) (Buscher et al. 2013). Additionally, for *T. brucei evansi* there is a rapid diagnostic test targeting a recombinant VSG produced in yeast (Birhanu et al. 2015). However, the VAP may

allow the development of diagnostics that not only screen for disease, but also predict its outcome, virulence and distribution. Epidemiological mapping based on variant antigens has been successfully achieved for other infectious diseases. In malaria, analysis of *var* gene diversity and expression patterns has revealed associations of particular genes to disease outcome. For example, there are *var* gene expression patterns characteristic from cerebral malaria (Kyriacou et al. 2006), whereas particular *var* gene groups have been associated with severe disease in Brazil and Africa (Kirchgatter & Del Portillo 2002; Chen et al. 2011; C. W. Wang et al. 2012). Therefore, if, as seems reasonable, the same happens for AAT, research on VSG expression may result in an increased understanding of clinical outcome.

The ability to compare and contrast VSG expression during infections and between isolates may also help vaccine design. Unlike *T. brucei*, the stability of the *T. vivax* VSG repertoire and the lack of recombination suggest that *T. vivax* could be amenable to vaccination. The literature reports that *T. vivax* infections can be 'self-limiting' in trypanotolerant cattle. Analyses of disease progression in experimentally infected cattle in Kenya suggested that the VSG repertoire of a single isolate could be exhausted, resulting in self-cure (Barry 1986). They also show cross-protection between isolates belonging to the same serodeme, suggesting that natural protective immunity is possible (Nantulya et al. 1986). Furthermore, in the early stages of *T. vivax* infection, the expressed VSGs of distinct isolates often belong to the same serodeme, perhaps reflecting some predictability of the expressed VSG sequences (Barry 1986). These results, coupled with my observation that variation among the *T. vivax* VSG repertoire is much lower than in *T. congolense* and *T. brucei*, offer the hope that there may be VSGs constitutively expressed in *T. vivax* infections, or else substantially lower antigenic diversity. As the VAP provides a way to study VSG diversity and expression patterns in *T. vivax*, it may help identifying potential VSG-based vaccine candidates for AAT.

5.4.2 The COG matrix methodology

The COG matrix was built using read mapping or sequence similarity searches. When doing read mapping with Bowtie2 (Langmead & Salzberg 2012), results are better under sensitive settings to account for strain diversity. The three most common DNA-seq aligners [i.e. Bowtie2 (Langmead & Salzberg 2012), BWA-Mem (Li 2013), and STAR (Dobin et al. 2013)] and the most recent segemehl (Otto et al.

2014) have similar accuracy rates (98.6-99.9 %), but Bowtie2 has the second to lowest user time and the lowest memory requirements (Otto et al. 2014). This is particularly important when building a method to handle multiple samples. Additionally, with the use of long-read genome sequencing increasing, these software are, or have been, updated to accept long reads, thus ensuring the adaptability of the VAP. This tool successfully avoids the need for annotation and genome assembly, which greatly enhances its speed and sensitivity, as genome assembly remains a significant problem for computational biologists (Henson et al. 2012).

It is difficult at this point to decide which screening approach is best. The sequence similarity approach is more stringent and is dependent on the assembly quality; therefore it finds fewer COGs in the dataset. However, since it relies on the similarity of the whole VSG contig, it is more accurate. The read mapping approach is more sensitive because, first, it bypasses genome assembly, and second, it is dependent on the read length rather than contig length. This results in more COGs being recovered, partly because the definition of COG becomes more relaxed. This can be partially overcome by increasing the genome read length, but will always be less strict than the comparison of two full VSG contigs. The presence of recombination is an issue that impacts both approaches, yet in different manners. While recombinant VSGs would not pass the rigorous requirements of the first approach, they might be considered in the second. In fact, in the latter, a mosaic VSG with two parents belonging to two distinct COGs would be seen as evidence for the presence of both parents, whilst in the former, it would be discarded and seen as a novel variant absent from the database.

Although strictness is often ideal, the ability to simultaneously deal with large fast-evolving gene families and incomplete genomes can become compromised. For example, if taking the most stringent parameters (contig search, 98 %), many COGs appear geographically-defined and the widespread VSG repertoire becomes as low as 5 % of the total VSG pool. Thus, the question becomes where to draw the line. In reality, it comes down to three aspects: the definition of ortholog, or how similar two VSGs must be to be considered true orthologs; whether VSGs are subject to the same degree of genetic variation as the rest of the genome; and how complete the genomes are. At this point, the solution is to invest in more sampling to produce an exhaustive COG database. In future, as the database becomes larger, it might be possible to refer to the *T. vivax* VSG repertoire as a finite number of individual

genes. In that case, as all expected variation is contained within the database, it would be easier to define orthologues and therefore identify the real location-specific VSGs.

5.4.3 Conclusion and Future directions

This work proposes the VAP as a new method to analyse antigenic diversity in *T. vivax* on a population scale and characterise VSG expression *in vivo*, overcoming the computational challenge of studying antigenic diversity on genomic and population scales. The particular nature of the *T. vivax* VSG repertoire revealed the importance of genome completeness and comprehensive sampling for an accurate VAP estimation. Despite the need for further sampling and deeper sequencing, this study showed that the *T. vivax* VSG repertoire is substantially more stable than that of *T. brucei* and *T. congolense*, raising the possibility of a finite number of antigens. However, I also identified a panel of location-specific VSGs that can be used as diagnostic markers in epidemiological studies. In future, we will link antigens to disease phenotypes, host susceptibility, and parasite virulence, allowing the VAP to deliver a better understanding of disease outcomes.

Chapter 6. The molecular evolution of VSGs in African trypanosomes

6.1 Introduction

In 2012, Jackson et al. showed that the VSG repertoires of *T. brucei*, *T. congolense*, and *T. vivax* produced quite distinct phylogenetic patterns, indicative of distinct underlying evolutionary mechanisms. By comparing the probability of phylogenetic incompatibility in the VSG repertoire of each species, they inferred that recombination played a variable role in diversity, since phylogenetic incompatibility probabilities are thought to reflect distinct recombination rates. In this chapter, I tested this hypothesis from a population perspective. By offering multiple viewing points of the same historical recombination event, a population analysis is more powerful than the examination of an individual reference genome for the historical signature of recombination past. Therefore, this chapter aims to understand the balance of evolutionary forces affecting VSG evolution in African trypanosomes, whether this is recombination, mutation rate, or population history.

Most of the *T. brucei* and *T. congolense* VSG genes are located in the subtelomeres, the regions between the interstitial regions of the chromosomes and their ends (Callejas et al. 2006). These regions are rich in species-specific genes and diverse sequence repeats due to high levels of transcription and recombination, and low levels of negative selection. In African trypanosomes, subtelomeres are larger and more variable than in other trypanosomatids probably due to the VSG system (Callejas et al. 2006). Here, recombination is thought to occur ectopically, is not restricted to homologous cross-over at cell division (Horn & Barry 2005) and there is affinity towards heterologous sequences, which greatly enhances gene diversification and expansion of gene families (Ricchetti et al. 2003). Pérez-Morga et al. (2001) also showed telomere clustering in *T. brucei* through *in-situ* hybridization analysis, which further facilitates recombination by bringing telomeres and subtelomeres into physical proximity, and makes subtelomeres the ideal regions for the VSG archive.

Homologous recombination also allows VSG mosaicism. VSG mosaic gene formation has long been recognised as a key aspect of infection chronicity in *T. brucei* (Kamper & Barbet 1992; Barbet & Kamper 1993). Gene conversion among VSG genes contributes to wide amplification of the available antigenic repertoire and increasing antigenic diversity. Homologous recombination affects *T. brucei* VSGs. However, species differences in sequence composition and variation suggest that *T. congolense* and *T. vivax* have evolved distinct mechanisms for generating antigenic diversity (Jackson et al. 2012). Despite antigenic variation being a phenotype common to all African trypanosomes, the genomic organisation and evolutionary mechanisms of the VSG repertoire seem to vary with the species and so might the role of homologous recombination.

While *T. congolense* VSG are structurally heterogeneous, *T. brucei* b-type VSG have passed through bottleneck evolution that resulted in the loss of all but one b-type VSG lineage, generating the single lineage phenotype observed today (Jackson et al. 2012). Phylogenetic analysis of the VSG repertoires of the three species shows clear differences in sequence organisation and variation. While *T. brucei* phylogenies are very “tree-like” with multiple long terminal nodes, consistent with frequent recombination, the phylogenies of *T. congolense* and *T. vivax* have longer basal nodes, creating long, separate clusters containing multiple, closely related sequences, structures that could only be maintained if recombinatorial pressures were low.

Evidence for recombination accounting for differences in VSG variation is strengthened by the frequency of VSG pseudogenes in the genomes. As pseudogenes are often the result of gene conversion and sequence reorganisation, their abundance in the genome is positively correlated with the frequency of recombination (Thon et al. 1989; Purandare & Patel 1997). Indeed, the number of pseudogenes in *T. brucei* is much higher than in *T. congolense* or *T. vivax* (69.2 % for a-type VSG, 72.2 % for b-type VSG, 21.1 % for Fam13, 29.7 % for Fam16, 15.5 % for Fam23, and 27.2 % for Fam24) (Jackson et al. 2012; Hall et al. 2013; Cross et al. 2014). The analysis of the phylogenetic incompatibility (PI) probability performed by Jackson et al. (2012) supports the same argument. PI is low amongst *T. vivax* VSG of the same family, regardless of their sequence identity (0.138 and 0.126 for Fam23 and Fam24, respectively), suggesting that recombination is scarce amongst this species VSG (Jackson et al. 2012). However, results for *T. congolense* support a different conclusion. Whilst phylogenetic incompatibilities are lower when

comparing random sequences of the VSG repertoire (0.125 for Fam13 and 0.43 for Fam16), they increase drastically when analysing sequences of the same phylotype (0.466 and 0.823 for Fam13 and Fam16, respectively) (Jackson et al. 2012). Furthermore, these probabilities can be increased if the CTDs are removed from the analysis, suggesting that recombination amongst *T. congolense* VSG occurs mostly within the same phylotype and that the CTD is a constraint for recombination (Jackson et al. 2012). The latter is a major difference with *T. brucei*, whose CTD is known to facilitate recombination by providing a sequence anchor point and containing a defined breakpoint. In fact, exchange of CTDs between VSG with distinct NTDs has been shown multiple times (Hutchinson et al. 2003). The analysis of PI in *T. brucei* VSG completes the argument by showing a generally high probability of recombination, but higher within the closely related sequences and a significant decrease when the CTD is removed (Jackson et al. 2012).

Homologous recombination is relevant for pathogenesis because it directly relates to antigenic switching. Antigenic switching requires the movement of a silent VSG from the subtelomeres or other telomeric expression sites to the active expression site, which is achieved by sequence recombination. However, homologous recombination is dependent on sequence homology and substrate length (Bell & McCulloch 2003; Barnes & McCulloch 2007), as well as the efficiency of base mismatch repair. *T. brucei* has at least three pathways of homologous recombination contributing to VSG recombination: the RAD51-dependent pathway, which has been well characterised in the literature and accounts for the majority of switching events (McCulloch & Barry 1999; Proudfoot & McCulloch 2005); RAD51-independent, microhomology-mediated end-joining (Glover et al. 2011); and the MSH2-independent pathway, which accounts for the residual VSG switching events observed in RAD51 mutants (Barnes & McCulloch 2007). Whilst the microhomology-mediated end-joining pathway is more efficient for substrates of 200 bp, the MSH2-independent pathway favours mostly short homology substrates of 5-15 bp (Barnes & McCulloch 2007).

As homologous recombination relies on DNA repair mechanisms, it has been proposed that DNA double-strand breaks (DSB) are triggers of VSG recombination (Glover et al. 2013). The frequency of natural DSB increases with the proximity to the telomere, suggesting that the subtelomeres are inherently fragile sites prone to recombination. However, not all DSB result in recombination and antigenic switching. Instead, the likelihood of a successful VSG recombination event is

dependent on the location of the DSB and the degree of homology of the flanking regions. Specifically, DSB occurring in the expression site downstream of the VSG and closer to the telomere are less efficient in triggering antigenic switching because the distance to the 70 bp repeat, which provides sequence homology to facilitate VSG recombination, is larger. Glover et al. (2013) presented a comprehensive model of antigenic variation that brings together all these aspects. In the event of a DSB at the expression site, which is frequent due to the fragile nature of the subtelomeres, the 70 bp repeats in the expression site associate with the 70 bp repeats of the subtelomeres to facilitate conversion of the active VSG as part of the DNA repair mechanism. Importantly, not all DSB occurring in the expression sites result in antigenic variation, which agrees with the experimental observation that antigenic switching occurs at much lower rates than subtelomeric DSB (Glover et al. 2013).

In this chapter, I directly compare the VSG repertoires of *T. brucei*, *T. congolense*, and *T. vivax*, at the population level, using a combination of read mapping and homologous recombination estimations. The specific aims are:

1. To give an interspecies perspective of VSG variability, which will allow VSG variability prediction for population genomic studies.
2. To quantify expected orthology among VSG repertoires to reveal the species-specific signature of gene hypervariability.
3. To investigate the role of recombination in generating VSG diversity across African trypanosome populations to understand what is the balance of evolutionary forces affecting VSG diversity.

6.2 Methods

6.2.1 Genomes

To evaluate VSG diversity on a population level, I have used twenty-six *T. congolense* genomes of those presented in chapter 2, all *T. vivax* genomes from chapter 5, twenty-one *T. brucei* and three *T. brucei evansi* genomes that were retrieved from the ENA (Table 16).

Table 16 Description of samples used in this chapter. Isolates are organised by species and include country of collection, host species, date of collection and previous publication details.

Isolate	Subspecies	Country	Host	Date collected	Publication
KP33 clone 16	<i>T. brucei brucei</i>	Ivory Coast	Tsetse fly	1989	Sistrom et al. 2014
LM 56 clone 6	<i>T. brucei brucei</i>	Ivory Coast	Pig	1992	Sistrom et al. 2014
KETRI1902	<i>T. brucei brucei</i>	Kenya	Waterbuck	1971	Sistrom et al. 2014
STIB213	<i>T. brucei brucei</i>	Tanzania	Hyena	1971	Sistrom et al. 2014
427 var 3	<i>T. brucei brucei</i>	Uganda	Sheep	1960	Sistrom et al. 2014
H884	<i>T. brucei brucei</i>	Uganda	Bovine	2003	Sistrom et al. 2014
TREU 927/4	<i>T. brucei brucei</i>	Kenya	Tsetse fly	1970	Gibson, 2012
MOS	<i>T. brucei gambiense</i>	Cameroon	Human	1974	Weir et al. 2016
YENB17_4	<i>T. brucei gambiense</i>	Guinea	Human	2002	Weir et al. 2016
DOB112_KIVI	<i>T. brucei gambiense</i>	Guinea	Human	2002	Weir et al. 2016
BROBT_16_KIVI	<i>T. brucei gambiense</i>	Guinea	Human	2002	Weir et al. 2016
S14_5_1	<i>T. brucei gambiense</i>	Ivory Coast	Human	2002	Weir et al. 2016
SETRA	<i>T. brucei gambiense</i>	Ivory Coast	Human	1979	Weir et al. 2016
KOBIR	<i>T. brucei gambiense</i>	Ivory Coast	Human	1982	Weir et al. 2016
ADZAM	<i>T. brucei gambiense</i>	Ivory Coast	Human	1983	Weir et al. 2016
SEVAL	<i>T. brucei gambiense</i>	Ivory Coast	Human	1984	Weir et al. 2016
B4_I314P	<i>T. brucei gambiense</i>	Ivory Coast	Human	2004	Weir et al. 2016
AMAN_KIVI	<i>T. brucei gambiense</i>	Ivory Coast	Human	2001	Weir et al. 2016
STIB809	<i>T. brucei rhodesiense</i>	Ethiopia	Human	1967	Sistrom et al. 2014
LVH56	<i>T. brucei rhodesiense</i>	Kenya	Human	1978	Sistrom et al. 2014
EATRO-0240	<i>T. brucei rhodesiense</i>	Uganda	Human	1961	Sistrom et al. 2014
H885	<i>T. brucei rhodesiense</i>	Uganda	human	2010	Sistrom et al. 2014
IL1180	<i>T. congolense</i>	Tanzania	Lion	1961	unpublished
IL1769	<i>T. congolense</i>	Tanzania	Lion	1971	unpublished
IL2281	<i>T. congolense</i>	Nigeria	Cattle	1979	unpublished
IL274	<i>T. congolense</i>	Kenya	Dog	1976	unpublished
IL2992	<i>T. congolense</i>	Kenya	Cattle	1966	unpublished
IL3019	<i>T. congolense</i>	Kenya	Cattle	1966	unpublished
IL3021	<i>T. congolense</i>	Kenya	Cattle	1966	unpublished
IL3022	<i>T. congolense</i>	Kenya	Cattle	1966	unpublished
IL3035	<i>T. congolense</i>	Kenya	Cattle	1985	unpublished
IL3180	<i>T. congolense</i>	Kenya	Cattle	1966	unpublished
IL3296	<i>T. congolense</i>	Tanzania	Cattle	1972	unpublished
IL3674	<i>T. congolense</i>	Gambia	Cattle	1979	unpublished
IL3675	<i>T. congolense</i>	Gambia	Cattle	1979	unpublished
IL3688	<i>T. congolense</i>	Tanzania	Lion	1961	unpublished
IL374	<i>T. congolense</i>	Kenya	Dog	1976	unpublished
IL3775	<i>T. congolense</i>	Kenya	Cattle	1966	unpublished
IL3897	<i>T. congolense</i>	Burkina Faso	Cattle	1982	unpublished
IL3949	<i>T. congolense</i>	Kenya	Cattle	1972	unpublished
IL3954	<i>T. congolense</i>	Nigeria	Cattle	1967	unpublished
IL396	<i>T. congolense</i>	Kenya	Dog	1976	unpublished
IL3978	<i>T. congolense</i>	unknown	unknown	1992	unpublished
IL399	<i>T. congolense</i>	Kenya	Dog	1976	unpublished
IL409	<i>T. congolense</i>	Kenya	Dog	1976	unpublished
IL410	<i>T. congolense</i>	Kenya	Dog	1976	unpublished
IL439	<i>T. congolense</i>	Kenya	Dog	1976	unpublished
ILC-22	<i>T. congolense</i>	Tanzania	Cattle	1970	unpublished
ILC-66	<i>T. congolense</i>	Kenya	Dog	1976	unpublished
E110	<i>T. evansi</i>	Brazil	Capibara	1985	Sistrom et al. 2014
STIB810	<i>T. evansi</i>	China	Buffalo	1985	Sistrom et al. 2014
KETRI 2479	<i>T. evansi</i>	Kenya	Camel	1980	Sistrom et al. 2014
IL340	<i>T. vivax</i>	Nigeria	Bovine	1962	unpublished
IL2323	<i>T. vivax</i>	Uganda	Tsetse fly	1969	unpublished
IL2005	<i>T. vivax</i>	Uganda	Tsetse fly	1969	unpublished
IL2714	<i>T. vivax</i>	Uganda	Tsetse fly	1969	unpublished
ILV-21	<i>T. vivax</i>	Uganda	Bovine	1972	unpublished
IL493	<i>T. vivax</i>	Nigeria	Bovine	1973	unpublished
IL11	<i>T. vivax</i>	Nigeria	Bovine	1973	unpublished
IL306	<i>T. vivax</i>	Nigeria	Bovine	1973	unpublished
IL462	<i>T. vivax</i>	Nigeria	Bovine	1973	unpublished
IL684	<i>T. vivax</i>	Nigeria	Bovine	1973	unpublished
IL465	<i>T. vivax</i>	Nigeria	Bovine	1973	unpublished
IL319	<i>T. vivax</i>	Nigeria	Bovine	1973	unpublished
IL338	<i>T. vivax</i>	Nigeria	Bovine	1973	unpublished
IL1392	<i>T. vivax</i>	Nigeria	Bovine	1981	unpublished
IL3658	<i>T. vivax</i>	Ivory Coast	Bovine	1990	unpublished
IL3638	<i>T. vivax</i>	Ivory Coast	Bovine	1990	unpublished
IL3651	<i>T. vivax</i>	Ivory Coast	Bovine	1990	unpublished
IL3171	<i>T. vivax</i>	The Gambia	Bovine	N/A	unpublished
IL596	<i>T. vivax</i>	Nigeria	Bovine	1973	unpublished

6.2.2 VSG mapping

To indirectly test the hypothesis that mosaicism rates are higher in *T. brucei* than the remaining species, I have compared VSG read mapping percentages across different strains. Specifically, VSG read mapping with Bowtie2 (Langmead & Salzberg 2012) was used to investigate the frequency of VSG conservation between the field strains and the reference. To retrieve VSG reads from the field strains, sequencing reads were mapped against a database of full length VSG and the number of mapped paired reads noted (**Figure 55**). The mapped reads were subsequently mapped against the database of full length VSG from the reference strain of each species and the number of mapped paired reads noted again (**Figure 55**). The percentage of reads from the field strain remaining paired in the reference strain was calculated and compared between species. The same procedure was applied to adenylate cyclases, which worked as a negative control for the baseline background variation (Alexandre et al. 1996; Bridges et al. 2008; Salmon et al. 2012).

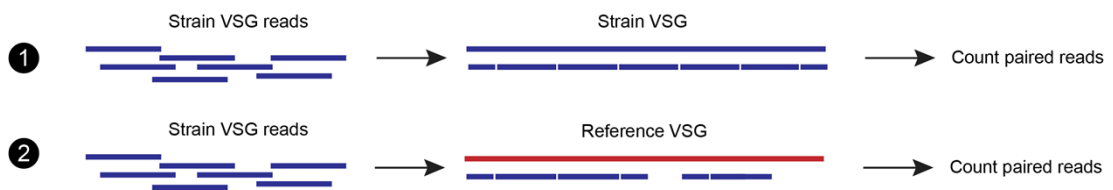


Figure 55 Reference VSG mapping strategy. Field strain VSG reads were retrieved by read mapping using Bowtie2 (Langmead & Salzberg 2012). These reads were mapped against a database of full-length VSG from the same strain and mapped paired reads were quantified (1). The same VSG reads were subsequently mapped to a database of full length VSGs from the reference genome (2). Mapped paired reads were quantified and compared to (1). This procedure was applied for all strains of *T. brucei*, *T. congolense*, and *T. vivax*

6.2.3 VSG characterisation

The variability in VSG conservation according to strain and species suggested that VSG mosaicism exists in different proportions between species. However, to understand how mosaicism is happening and the players involved, an approach inverse to the first one was used. Rather than using field strain reads, reference

VSG were fragmented into 150 nucleotide pseudo-reads and mapped against the full-length strain VSG (**Figure 56**). This was necessary because the field strain VSG repertoires are incomplete and of variable sizes, which would introduce greater variability and error when quantifying mapped and unmapped reads.

The original identifiers of the reference VSGs were kept in the pseudo-reads so that the frequency of reference reads remaining paired in the field VSG could be estimated. This procedure was applied to all strains and compared between species. VSG mapping was performed with Bowtie2 under sensitive settings (Langmead & Salzberg 2012). Using the mapping results, all VSG from the field strains were characterised into uncoupled, multi-coupled and fully coupled, according to how many reference VSG contributed to them (**Figure 56**). The reference VSGs contributing to a strain VSG were called 'donors'. Fully coupled VSG were those with at least one donor contributing to more than 84 % of the sequence. Multi-coupled VSG were those with one or more donors contributing with more than 1 pseudo-read (≥ 300 nucleotides), whereas uncoupled VSG were those remaining [i.e. one or more donors contributing with 1 read only (i.e. ≤ 150 nucleotides)]. The reference VSG that were not mapped at least once to the strain VSG were considered reference-specific variants. Under these definitions, I would expect that *T. vivax* presented the highest percentage of fully coupled VSG, and *T. brucei* the highest percentages of multi-coupled and unmapped VSG.

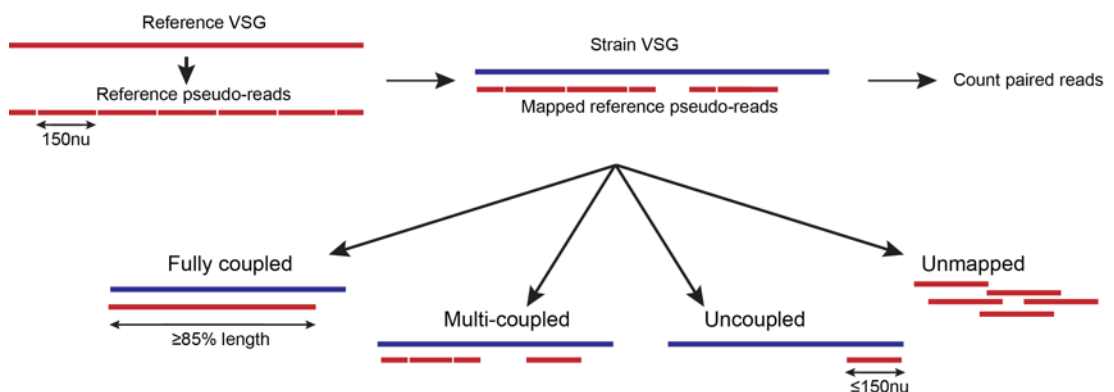


Figure 56 VSG characterisation strategy. Reference VSG sequences were split into pseudo-reads of 150 nucleotides. These pseudo-reads were mapped against a database of field strain full-length VSGs. Each field strain VSG was then defined as fully coupled, multi-coupled, uncoupled, or unmapped, according to how many and how much of the VSG pseudo-reads were mapped to it.

6.2.4 VSG donor analysis

To further understand how mosaicism is happening, the reference sequences contributing to each multi-coupled VSG (i.e. donors) were retrieved and analysed (**Figure 57**). For each VSG, the number of donors and their coverage (i.e. how many base pairs each donor contributed to the VSG) were calculated from the mapping output using a customised script. This approach allowed me to test the hypothesis that the level of mosaicism is lower for *T. vivax*, even within mosaic VSGs.

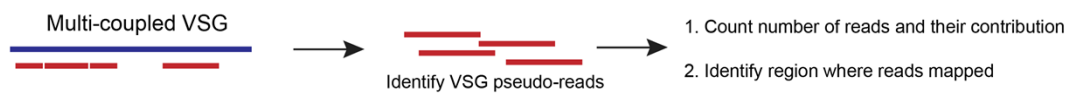


Figure 57 VSG donor analysis strategy. The VSG pseudo-reads mapped to multi-coupled VSG were retrieved and analysed. The number and contribution of pseudo-reads mapping to each VSG were quantified.

Subsequently, I performed a specific VSG donor analysis for *T. congolense* to understand whether closely related VSG sequences were more likely to recombine. The CTD is an anchor for recombination in *T. brucei* and has been proposed to have the same role in *T. congolense*. As each phylotype has a different CTD, intra-phylotype recombination has been shown to be the most frequent in the reference strain IL3000. To test whether the same was observed at the population level, *T. congolense* multi-coupled VSGs and their reference donors were profiled using the Variant Antigen Profiling tool presented in chapter 2. Intra- and inter-phylotype recombination were calculated and compared on a data matrix.

I have also tested whether particular regions of the *T. congolense* VSG sequence were more likely to be shared across the sample cohort. In chapter 2, I showed that the *T. congolense* 15 VSG phylotypes were mainly defined by conserved CTDs. This suggests that *T. congolense* uncoupled VSGs would most likely be mapped at the CTD. To test this hypothesis, the reference pseudo-reads mapping to *T. congolense* uncoupled VSGs were localised in the VSG secondary structure using their sequence identifier.

6.2.5 Recombination analysis

Whilst all the previous methodological approaches have measured recombination indirectly, the likelihood of recombination can be directly quantified by measuring PI. Alignments with recombinant sequences have multiple phylogenetic signals, i.e. different PI, whose probability can be estimated using the pair-wise homoplasy index (PHI) (Bruen et al. 2006).

To test whether multi-coupled VSGs were more likely to be recombinant than fully coupled VSGs, I calculated PI for each VSG group. PI was also calculated for adenylate cyclases as a control for the background genome recombination level and for two sets of simulated data (250 replicates, 16 sequences per replicate) with and without recombination. Simulated data was generated with NetRecodon (Arenas & Posada 2010), under diploid settings, a population mutation rate (θ) of 160, a heterogeneity rate of 0.05, and an expected population size of 1000. The population recombination rate (ρ) was set to 0 and 96 for the non-recombinant dataset and recombinant datasets, respectively. Both experimental and simulated sequences were divided into subsets of 4 sequences (quartets) and aligned using clustalW (Larkin et al. 2007). PHI was run iteratively (Bruen et al. 2006) on each quartet and the probability of PI recorded. The percentage of quartets with significant PI was compared between sequence types and species. The multi-coupled VSG group resulted in 909, 513, and 878 quartets for *T. brucei*, *T. congolense*, and *T. vivax*, respectively. The fully coupled VSG produced 115, 25, and 159 quartets, and the adenylate cyclases group yielded 910, 155, and 269 quartets, respectively. The simulated data resulted in 1000 quartets per condition.

6.3 Results

6.3.1 VSG mapping

To indirectly evaluate how much of the VSG repertoire was mosaic in the population, I quantified the proportion of VSG reads from the different strains remaining paired in the reference full repertoire. To test the hypothesis that the different species had different degrees of VSG repertoire conservation, I calculated the percentage of paired reads shared between the reference and the field strains. This showed that *T. brucei* had the lowest percentage of shared paired reads (79 %), whilst *T. congolense* and *T. vivax* had 87 % and 94 %, respectively (**Figure 58**). The same approach was used with adenylate cyclases as a negative control for the background genetic variation. In contrast with VSG, the percentage of adenylate cyclases shared paired reads was 94 %, 98 %, and 97 % for *T. brucei*, *T. congolense* and *T. vivax*. For each species, the difference in shared paired reads between VSG and adenylate cyclases was significantly different (Independent t-test, $p < 0.001$). Within the VSG group, the species differences are also significant (**Figure 58**).

Furthermore, **Figure 58** shows that the averages presented above are the result of very distinct dynamics. *T. vivax* shows a pattern where the percentage of shared reads between the strains and the reference is always higher than 90 %. The majority of *T. brucei* only share between 70 and 79 %, whilst a minority shares 80-89 %, and *T. congolense* presents an intermediate scenario where the percentage of shared reads increases from 60-69 % (N = 2 strains) to more than 89 % (N = 12). A lower percentage of shared paired reads indicates that, in different strains, VSG reads map to distinct regions, which is a likely result of lower conservation between VSGs. Therefore, these data suggest that *T. brucei* has the most variable VSG repertoire and *T. vivax* occupies the opposite end of the spectrum with the highest degree of conservation among strains.

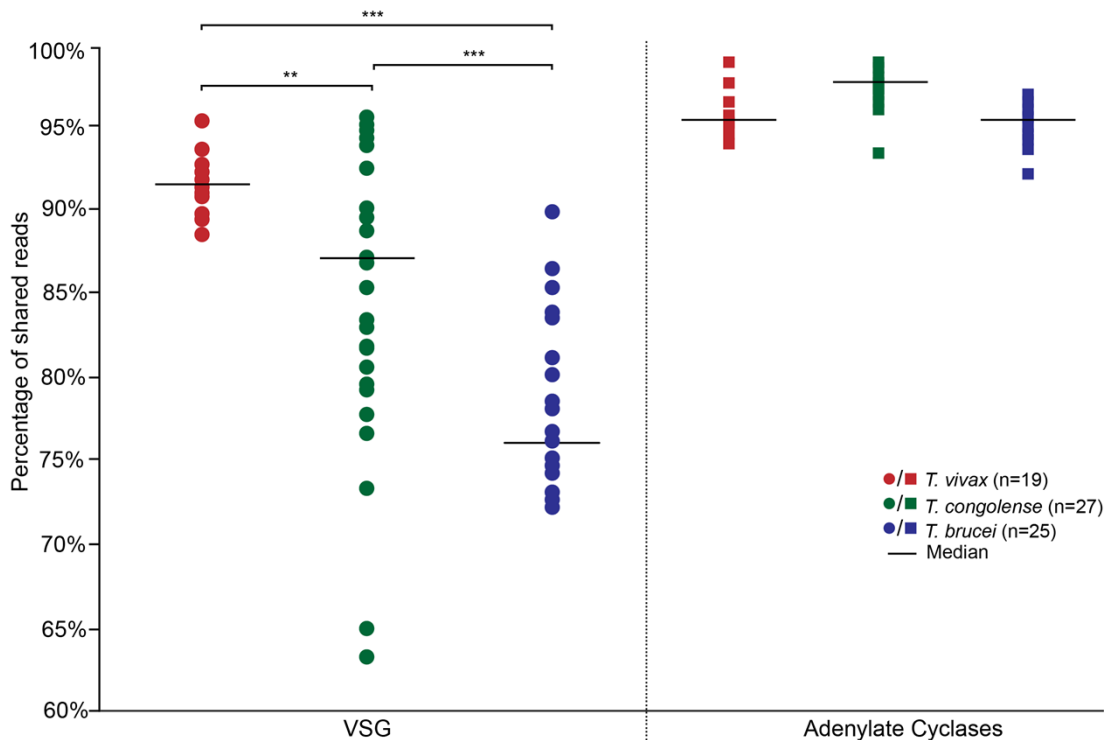


Figure 58 Proportion of field strain paired VSG reads remaining paired when mapped to full-length VSG for each African trypanosome species. Adenylate cyclases (AC) were included as a negative control for background mapping variation. African trypanosome species are colour-coded according to key. Stars indicate statistical significant differences between species (Independent t-test, ** = $p < 0.01$; *** = $p < 0.001$). The percentage of shared reads between VSGs and adenylate cyclases was significantly different for all species (Independent t-test, $p < 0.001$).

6.3.2 VSG characterisation

Analysis of read mapping indicates that species differ in VSG rearrangement, but is not sensitive enough to show the scale of mosaicism. To achieve this, the reference VSGs were segmented into non-overlapping pseudo reads of 150 nucleotides and iteratively mapped against each strain full-length VSGs. The reference VSG pseudo reads were named based on their original VSG and their positioning in the original sequence so they could be tracked after mapping. The mapping outputs were used to calculate read coverage and to classify strain VSGs accordingly into three groups: fully coupled, or those strain VSGs whose reference VSG pseudo reads aligned to more than 84 % of sequences; uncoupled, or those strain VSGs to which only one pseudo read per reference VSG mapped; and multi-coupled, or those

strain VSGs which have had multiple pseudo reads from the same reference VSGs mapped to them, but which cover less than 85 % of the full-length VSG sequence (**Figure 59**). Additionally, those reference VSGs that did not map to any of the strain VSGs were considered unmapped and therefore unique to the reference in the context of the population being analysed.

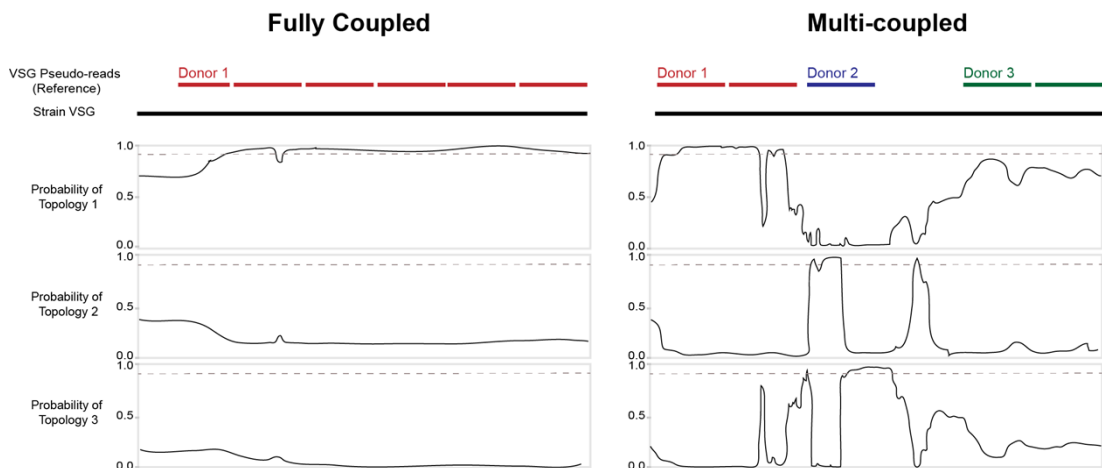


Figure 59 Example representation of fully coupled and multi-coupled VSGs with their respective topology probability. Topologies were estimated with Topali v2 (Milne et al. 2009). Breaks in topology significance reflect evidence for recombination. In fully coupled sequences, only one topology is significant across the whole nucleotide sequence. In contrast, in a multi-coupled sequence, three different topologies account for different parts of the nucleotide sequence, reflecting at least three VSG donors. Dashed line shows significance threshold of 95 %.

This approach showed that *T. vivax* has a significantly higher percentage of fully coupled VSGs than both *T. congolense* and *T. brucei* (59 ± 4 %, Independent t-test, $p < 0.0001$). In contrast, the percentages of fully coupled VSGs in *T. brucei* and *T. congolense* are not significantly different. *T. congolense* has similar percentages of multi-coupled and uncoupled VSGs (33 ± 2 % and 31 ± 4 %). The percentage of *T. congolense* multi-coupled VSGs is significantly higher than that of *T. vivax* (Independent t-test, $p < 0.01$), but significantly lower than *T. brucei* (Independent t-test, $p < 0.05$). Therefore, *T. brucei* has the highest percentage of multi-coupled VSGs (39 ± 1 %) and *T. vivax* the lowest (25 ± 2 %). In contrast, the percentage of *T. congolense* uncoupled VSG is significantly higher than both *T. vivax* (8 ± 1 %) and *T. brucei* (12 ± 1 %) (Independent t-test, $p < 0.001$).

A high percentage of fully coupled VSGs suggests high repertoire conservation in the population. In contrast, multi-coupled and uncoupled VSGs are mosaics. These results suggest higher VSG conservation among *T. vivax* strains. Hence, this confirms the hypothesis that *T. vivax* has the most conserved VSG repertoire among the sample cohort.

T. brucei has a significantly higher percentage of unmapped VSGs (29 ± 2 %) than both *T. vivax* (8 ± 2 %) and *T. congolense* (16 ± 1 %) (Independent t-test, $p < 0.0001$) (**Figure 60**). Unmapped VSGs reflect strain-specific VSGs, which can result from gene turnover high enough to erase evidence of ancestral sequences. Therefore, the results also suggest that *T. brucei* VSG repertoire is the most variable across strains.

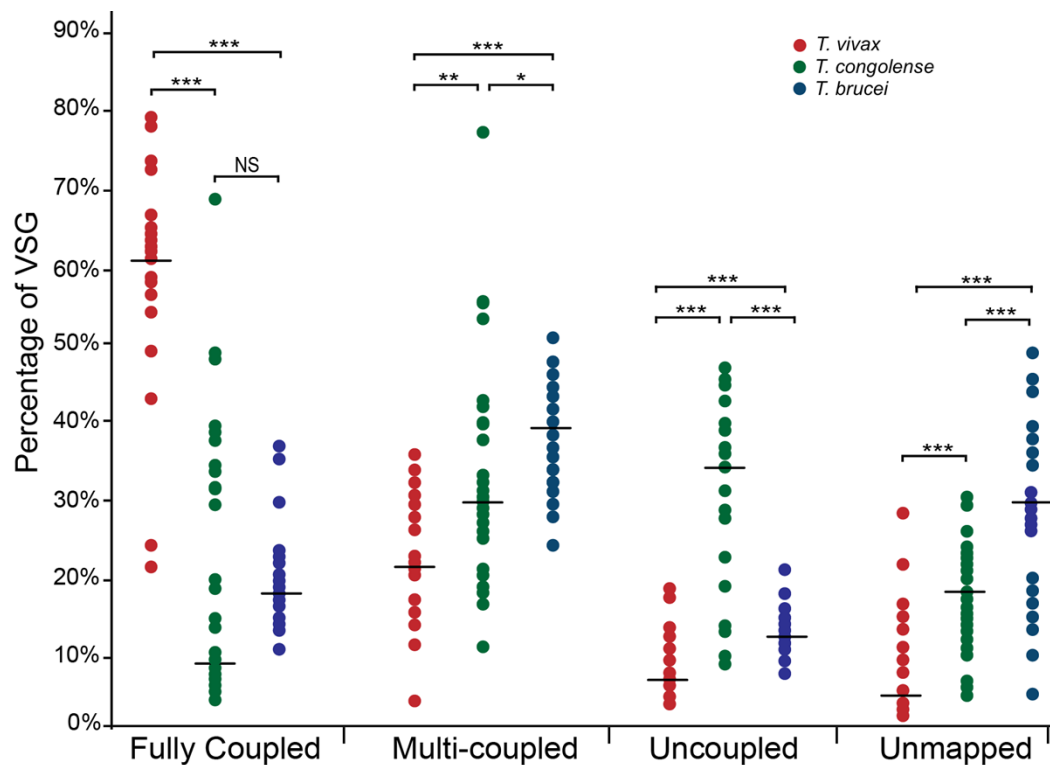


Figure 60 The composition of the VSG repertoire for each African trypanosome species (mean \pm SEM). Field strain VSGs were characterised according to the number of reference pseudo-reads that mapped to them into fully coupled, multi-coupled, uncoupled, or unmapped. Graph represents the percentage of each category per strain, per species. Stars indicate statistically significant differences (Independent t-test, * = $p < 0.05$; ** = $p < 0.01$; *** = $p < 0.001$). Species are colour-coded according to key.

Genomic novelty is mostly generated by recombination. In fact, in *T. brucei*, segmental gene conversion has already been showed to be crucial for new VSG variant generation (Marcello & Barry 2007a). So far read mapping results show that multi-coupled VSGs are mosaics. However, to test whether mosaicism is caused by recombination, I compared the probability of PI in fully coupled and multi-coupled sequences for each species. PI was calculated with PHI (Bruen et al. 2006) (**Figure 61**). This measure provided a statistical basis to link multi-coupling with recombination and thus validates the mapping-based methodology used in this chapter. The majority of multi-coupled VSG from all three species showed evidence for recombination (67 %, 65 %, and 41 %, for *T. brucei*, *T. congolense*, and *T. vivax*, respectively), whilst the evidence for recombination within the fully coupled VSGs is within the same level of both the background recombination level expressed by the arbitrarily selected adenylate cyclases quartets and the negative recombination simulations.

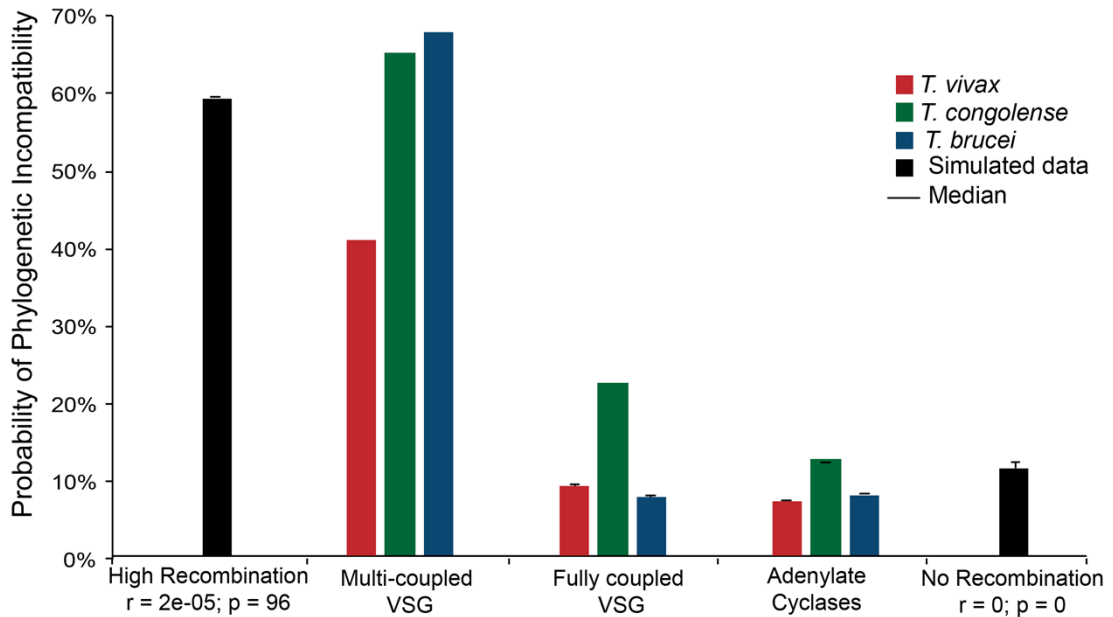


Figure 61 Phylogenetic incompatibility among VSG genes using PhiPack (Bruen et al. 2006). The percentage of sub-alignments showing significant phylogenetic incompatibility (Ppi) in in multi-coupled and fully coupled VSG is shown for *T. vivax* [N(MC) = 878; N(FC) = 444], *T. congolense* [N(MC) = 513; N(FC) = 213], and *T. brucei* [N(MC) = 909; N(FC) = 597], with bars shaded by species according to key. Adenylate cyclases were included as the experimental negative control [N(*T. vivax*) = 269 , N(*T. congolense*) = 155, N(*T. brucei*) = 910]. The same metric for 1000 quartets of simulated data in the presence and absence of recombination, obtained with NetRecodon (Arenas & Posada 2010), is shown in black. The population mutation rate (θ) was kept constant at 160; r represents the population recombination rate used to produce the sequences. African trypanosome species are colour-coded according to key.

The high percentage of uncoupled VSGs in *T. congolense* shown in **Figure 60** was surprising unless they represented alignments to the well-conserved CTDs. The *T. congolense* VSG repertoire is divided into 15 universal and exhaustive phylotypes characterised by distinctive conserved CTDs and therefore such a high percentage of uncoupled VSG should reflect this. This hypothesis was tested by tracking the reference pseudo-reads in the mapping output of each strain and localising the region of the strain VSG they matched. The results presented in **Figure 62** corroborate the predictions, showing that, in 58 % of the events, the reference pseudo read mapped to the CTD (**Figure 62**).

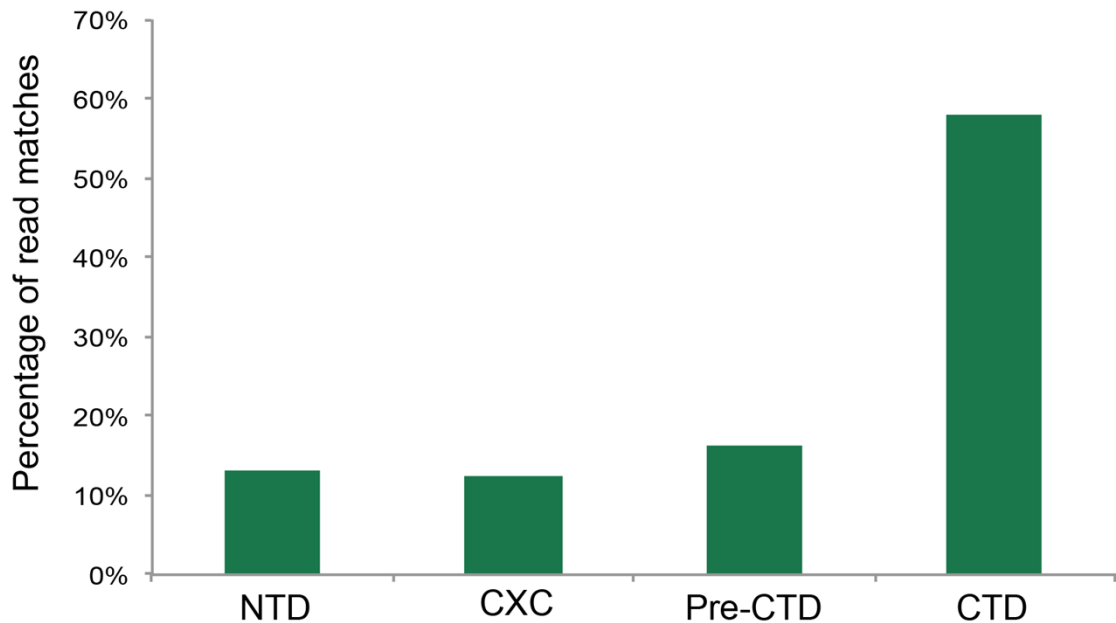


Figure 62 Donor VSG sequence localisation within *T. congolense* uncoupled VSGs.

NTD = N-terminal domain, CXC = the central CXC VSG domain, CTD = C-terminal domain, Pre CTD = the regions immediately upstream the CTD.

Another unusual observation was that, on average, 8 % of the *T. vivax* reference VSG pseudo reads remained unmapped (**Figure 60**). If the *T. vivax* VSG repertoire is as conserved as suggested by chapter 5 and **Figure 58**, then all reads should map to all strains. However, as shown in chapter 5, the *T. vivax* VSG repertoire contains a subset of region-specific VSG that are not found in all strains. Therefore, and given that the reference strain is Nigerian, it would be expected that the majority (if not all) of the unmapped reads should come from the non-Nigerian strains. To test this, the percentage of unmapped reads per strains was analysed. This revealed that among the Nigerian isolates the average percentage of unmapped reads was only $2\% \pm 1.44$, whilst the for non-Nigerian strains this increases to $16\% \pm 6.94$. These results suggest that whilst the majority of *T. vivax* VSGs are conserved across isolates, there is a small number of location-specific VSGs. These findings corroborate those presented in chapter 5.

6.3.3 Donor analysis

To further dissect the differences in VSG repertoire for the three African trypanosome species and to understand the reasons and mechanisms behind such differences, the number and contribution of the reference VSG ('donor') to each strain VSG was analysed. A lower number of donors per VSG would indicate that

not all VSG were free to recombine with each other, suggesting more constraints for VSG recombination. Although the three species have VSG with multiple donors, *T. brucei* has the highest number (up to 58), whereas *T. vivax* has the lowest (up to 29). These results support previous observations that mosaicism plays a large role in antigenic variation in *T. brucei* (Marcello & Barry 2007a; Hall et al. 2013). In fact, it has been shown to accumulate rapidly, contributing to a large sequence turnover, where sequence identity does not seem to constrain gene conversion (Hall et al. 2013). Our data also supports this view: we see that the total nucleotide identity between donors and their recipient VSG varies (**Figure 63**). Whilst the *T. vivax* donors share $86 \% \pm 0.1$ sequence similarity, the *T. brucei* donors and their correspondent VSG only share $47 \% \pm 0.06$, which indicates a higher repertoire conservation and greater orthology amongst VSGs in *T. vivax*, and consequently higher sequence diversity and fewer recombination constraints in *T. brucei*.

Whilst the number of donors in certain *T. vivax* VSGs is higher than expected for such recombination scarcity, it reflects the repetitive nature of the repertoire, populated by multiple VSG tandem arrays originated by duplicative gene conversion events. On the other hand, in *T. congolense*, the sequence identity required for recombination is high, closer to *T. vivax* at $71 \% \pm 0.13$, which likely reflects the recombination constraints between different VSG phylotypes. In fact, 88 % of *T. congolense* multi-coupled VSGs have donors of the same phylotype. This is shown in **Figure 64**, a heat map representing the relative frequency of phylotype combinations between the reference donor VSG and the strain VSG, confirming the hypothesis that recombination between sequences of different phylotypes is rare (Jackson et al. 2012). Further to this analysis, I looked at the contribution of each individual donor to the VSG in terms of sequence coverage and show that despite the large range (i.e. 8 to 84 %), lower coverages are more frequent in *T. congolense* and *T. brucei* than in *T. vivax* (**Figure 65**), suggesting a clear difference in repertoire conservation across the species among the three African trypanosomes.

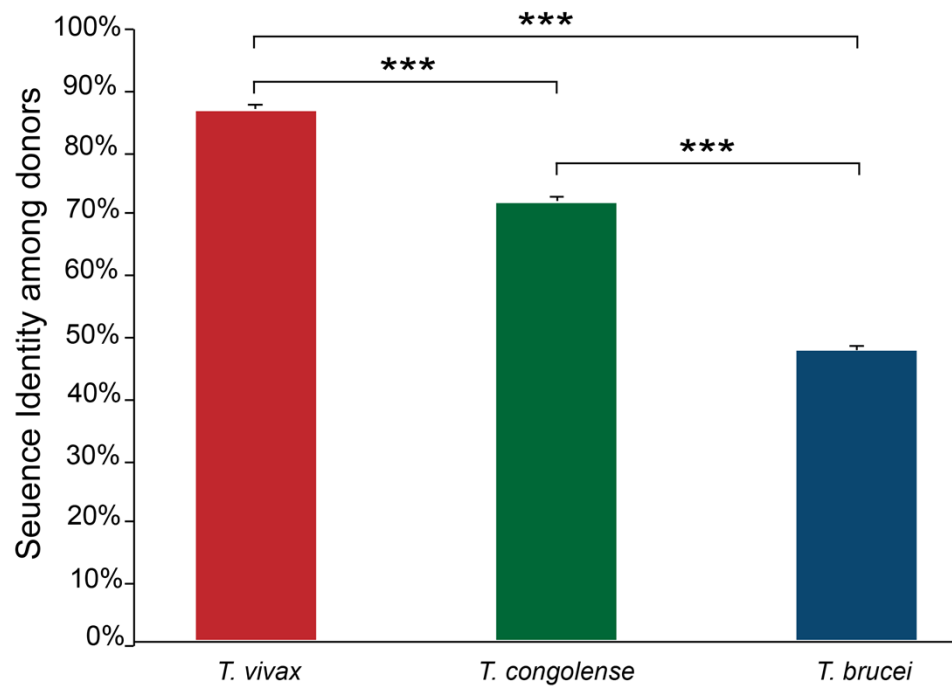


Figure 63 Nucleotide sequence identity amongst all sequences involved in MC VSG formation (donors and recipients) (mean \pm 95 % CI). Stars indicate statistical significant differences (Independent t-test, * = $p < 0.001$). African trypanosome species are colour-coded as previously.**

		Donor Phylotype														
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
VSG Phylotype	1	0.88	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.07	0.00	0.00
	2	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	3	0.05	0.03	0.92	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	4	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	5	0.00	0.00	0.00	0.09	0.76	0.12	0.01	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	6	0.00	0.00	0.00	0.14	0.07	0.69	0.01	0.09	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	7	0.00	0.00	0.00	0.00	0.00	0.00	0.99	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	8	0.00	0.00	0.00	0.33	0.27	0.00	0.00	0.40	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.93	0.00	0.07	0.00	0.00	0.00	0.00
	10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.85	0.00	0.00	0.00	0.00	0.15
	11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.88	0.08	0.02	0.01	0.00
	12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.15	0.85	0.00	0.00	0.00
	13	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.01	0.96	0.00	0.01
	14	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.69	0.31
	15	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.97

Figure 64 *T. congolense* VSG recombination distribution by phylotype. Data is expressed as a proportion of total recombination and colour-coded by number of events. Green represents lower and red represents higher frequency.

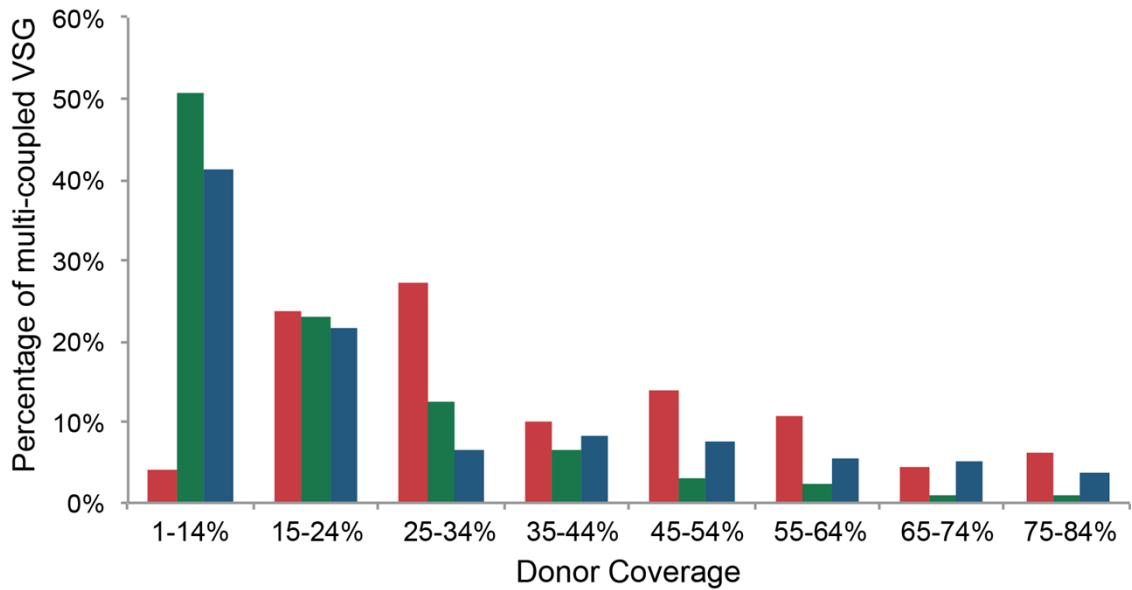


Figure 65 The sequence coverage of VSG donors expressed as a percentage of the full multi-coupled VSG repertoire. African trypanosome species are colour-coded as previously.

In all three species, the number of VSGs with a single donor was remarkably high. When these events are analysed further, they reveal that the majority of *T. vivax* single donors account for over 75 % of the sequence. This contrasts with *T. brucei* and *T. congolense*, where those instances are a minority, once again suggesting a clear difference in repertoire conservation across the species among the three African trypanosomes (**Figure 66**). The relative frequency of single donor coverage in *T. congolense* and *T. brucei* follows a normal distribution. The *T. congolense* maximum frequency donor coverage is between 25 and 44 %, and in *T. brucei* is 35 to 54 %.

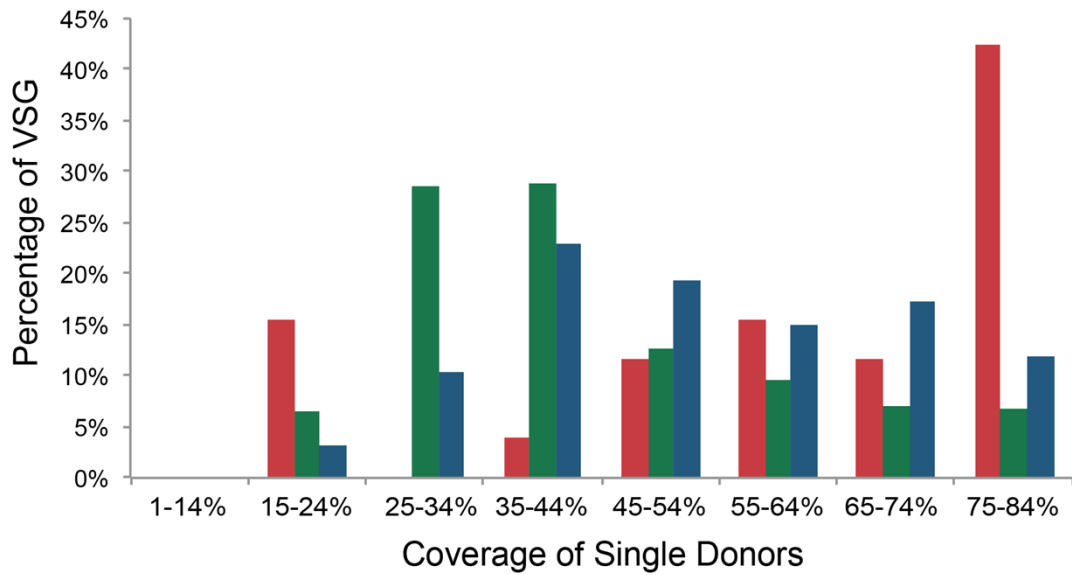


Figure 66 The sequence coverage of VSGs with a single donor, represented as a percentage of VSG. African trypanosome species are colour-coded as previously.

What remains unresolved is the degree of conservation that can be expected in the VSG repertoire across the population for each trypanosome species. In principle, orthology can be easily extrapolated from the percentage of fully coupled VSGs observed in each species ($59\% \pm 4$; $22\% \pm 4$; 20 ± 1 , for *T. vivax*, *T. congolense*, and *T. brucei*, respectively). Here, we have used the reference genomes as fixed points for pairwise comparisons with the remaining strains. If such analysis were repeated using a distinct genome as reference, it is highly likely that the specific set of FC VSG would change. As such, it is our view that the majority, if not all, VSGs are mosaic sequences, depending on which strains are being compared. Therefore, estimating orthology based exclusively on the set of FC VSG would result in a large underestimation of sequence conservation and in the overall interpretation that VSG orthology across the population is scarce. Thus, we propose that a more accurate way to calculate orthology amongst VSGs is to include the orthologous VSG sub-sequences that constitute mosaic VSGs and assess the number of shared base pairs in the context of the full VSG repertoire of each strain. This way, sequence orthology in the *T. brucei* VSG population repertoire can be estimated at 40 %, in *T. congolense* at 41 % and in *T. vivax* at 74 % (**Figure 67**).

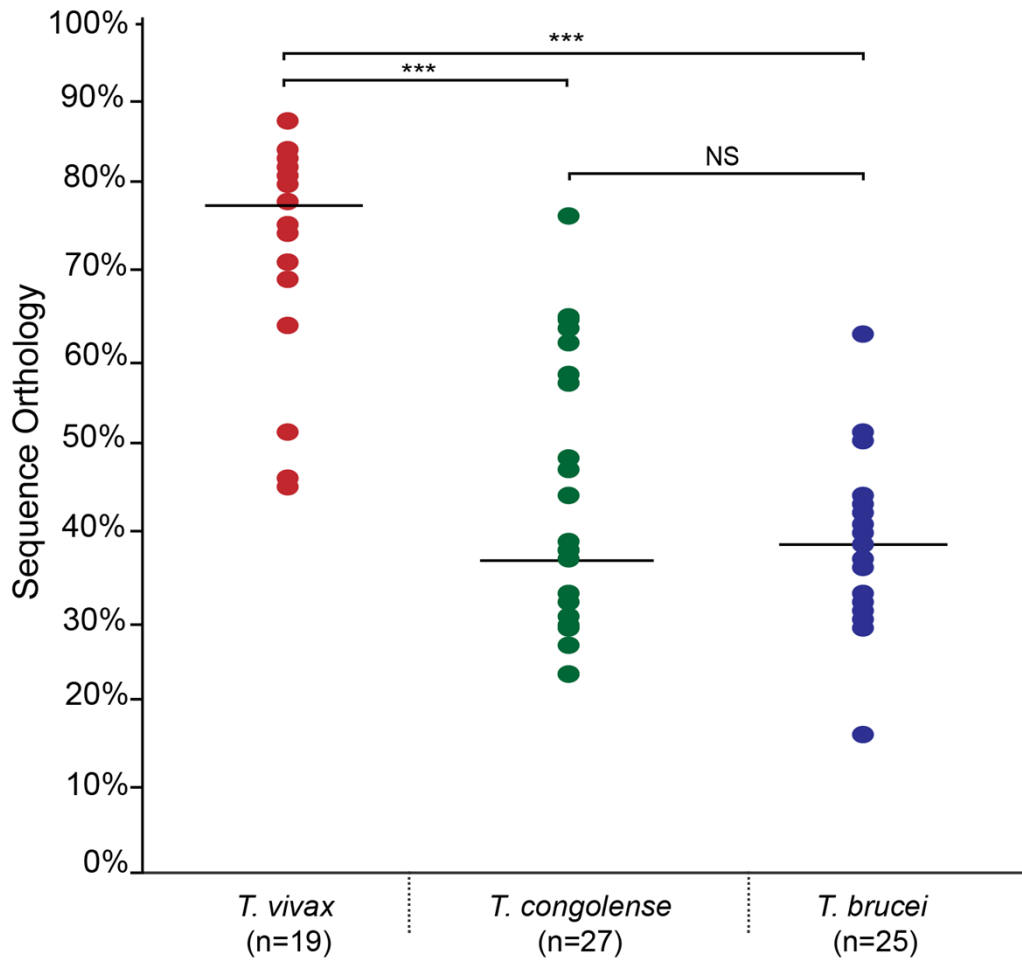


Figure 67 Total sequence orthology amongst VSG repertoires of the same species. Orthology was calculated as the proportion of shared nucleotides between each strain and the reference. Stars indicate statistical significant differences (Independent t-test, NS = non-significant; *** = $p < 0.001$). African trypanosome species are colour-coded as previously.

6.4 Discussion

Gene conversion has been previously proposed to play distinct roles in diversity generation in African trypanosomes (Jackson et al. 2012), but this hypothesis was based exclusively on the reference genomes of *T. brucei*, *T. vivax*, and *T. congolense*. In this study, I have analysed the conservation of the VSG repertoire across a population sample (N = 69) with a continental distribution. This has allowed me to compare the quantitative contribution of recombination to VSG diversity among the different African trypanosome species. I have shown that the rate of gene conversion in the VSG repertoire is highest among *T. brucei* isolates and lowest in *T. vivax*, whilst *T. congolense* occupies an intermediate position characterised by frequent recombination within specific gene cohorts. The novelty of this work lies in the overall picture of the role of recombination in sequence evolution afforded by population analysis, which is more accurate than the analysis of single isolates, especially when dealing with variant gene families undergoing extensive gene turnover.

Segmental gene conversion, or VSG mosaicism, has the main function of generating novel variants, utilising the vast repertoire of silent and/or pseudogenic VSG (Marcello & Barry 2007a). Furthermore, mosaicism allows infection chronicity by providing an endless pool of variant antigens. In *Anaplasma marginale*, mosaicism has been proposed as driver of super infections of partially immune hosts (Brayton et al. 2002). If that applies to African trypanosomes, then mosaicism adds complexity to the already vast phenotypic potential of trypanosomes (Barry et al. 2005).

The accepted hypothesis for the formation of mosaic VSGs in *T. brucei* is a pathway of sequential, independent steps of segmental conversion, enabled by a restricted set of donor VSG sequences (Marcello & Barry 2007a). This is consistent with sequence homology between sequence donor and recipient being a major determinant of recombination (Bell & McCulloch 2003). The *T. brucei* VSG repertoire is very rich in mosaic VSGs. For instance, unique VSG comprise 60 % of the *T. brucei* TREU927 VSG repertoire; the remaining 40 % relate to clusters of orthologues of 2-4 members with over 50 % identity (Marcello & Barry 2007a). In this chapter, I have shown that 50 % of *T. brucei* VSGs are mosaics (i.e. 39 % MC and 11 % UC) (**Figure 60**), of which 67 % have a recombination signature measurable by PI (**Figure 61**). These estimations are likely to be under-represented, because 29 %

of *T. brucei* VSGs lack any orthology within the population sample (i.e. the unmapped VSGs) and therefore were not analysed for evidence of recombination. These data suggest that the *T. brucei* VSG repertoire is highly dynamic and heavily influenced by ectopic recombination, which actively contributes to VSG diversity generation within the population.

Recombination is also a major generator of antigenic diversity in other organisms. In *P. falciparum*, the chromosomal ends are populated by *var* genes. Like in *T. brucei*, these telomere-associated structures are well conserved across chromosomes, favouring sequence recombination. Indeed, ectopic recombination is a major driver of new *var* genes formation (Freitas-Junior et al. 2000; Taylor et al. 2000). Likewise, in *Anaplasma marginale*, *msp2* gene expression is often preceded by multiple levels of recombination to convert pseudogene segments into a fully functional gene at the expression site (Brayton et al. 2002).

The results of this chapter indicate that recombination within the *T. congolense* VSG repertoire is dependent on sequence diversity. The analysis of *T. congolense* phylotype recombination showed that sequence exchange between sequences of different phylotypes is rare. The majority of VSG recombination events in *T. congolense* occur between sequences of the same phylotype, most likely due sequence homology constraints: the large disparities in sequence composition between Fam13 and Fam16 and amongst the CTDs of each phylotype appear to limit recombination. This agrees with the previous analysis of the IL3000 VSG repertoire (Jackson et al. 2012) and also with evidence presented in chapter 2 that the different phylotypes are under purifying selection. In *T. brucei*, such constraints do not exist, as the VSG CTDs are homogeneous, apparently facilitating homologous recombination between diverse sequences (Marcello & Barry 2007a; Jackson et al. 2012).

The presence of recombination constraints is even more evident in *T. vivax* than in *T. congolense*. The high number of fully coupled genes across the species and the scarce evidence of recombination are all indicators of high repertoire conservation. Yet, *T. vivax* seems to have the tools necessary for recombination, at least regarding the processes linked to sequence repair in *T. brucei*, such as the RAD51-dependent pathway double-strand breakpoint repair mechanism (for example, the DNA repair protein RAD51, the meiotic recombination protein DMC1, and the replication factor A protein RPA1), and the DNA mismatch repair protein MSH2 (Bell & McCulloch 2003;

Barnes & McCulloch 2007; Glover et al. 2011; Glover et al. 2013). Therefore, the low levels of recombination observed in *T. vivax* do not seem to relate to any loss in sequence repair ability. Although it is possible that these mechanisms are less efficient in *T. vivax*, therefore resulting in lower recombination levels, it is perhaps more likely that recombination constraints relate to differences in sequence homology as observed for *T. congolense*. As described in chapter 5, the *T. vivax* VSG repertoire is composed of 96 phylotypes. These lineages are separated by long evolutionary distances, but composed of a small number of very similar genes. As sequence homology strongly affects efficiency of ectopic recombination, it is possible that recombination is rare amongst *T. vivax* because the number of genes between which conversion is possible is limited and so is the number of pseudogenes. In fact, in this species the sequences involved in mosaic VSG formation usually share high nucleotide sequence identity (multi-coupled VSG, 86 % \pm 0.1), whilst *T. brucei* sequences are only 47 % \pm 0.06 similar (**Figure 63**).

If sequence homology is the major recombination constraint affecting *T. congolense* and *T. vivax*, it brings back the question of why and when the different lineages were formed. The lineages of *T. congolense* were likely present in the *T. congolense*/*T. brucei* ancestor because *T. brucei* ESAG2 closest relatives are members of phylotype 3 in *T. congolense*. Yet, these lineages are not in *T. vivax*, suggesting that they were developed after *T. vivax* speciation. This leaves the origin of the many *T. vivax* lineages unsolved. VSG families Fam23 and Fam24 are a- and b-type VSG, hence they were present before *T. vivax* speciation and have shared origin with *T. brucei* and *T. congolense* VSGs (Jackson et al. 2012). However, Fam25 and Fam26 are apparently *T. vivax*-specific (Jackson et al. 2012). They could be an innovation, which has arisen after the origin of *T. vivax*, or they could have been present in the African trypanosome ancestor and lost by the *T. congolense*/*T. brucei* ancestor.

In other antigenically variant organisms, recombination constraints within variant antigen gene repertoires are less evident. An analysis of *P. falciparum* var gene diversity from three distinct isolates has demonstrated that parasite isolates share little of their var gene repertoire, a sign of extensive segmental gene conversion (Kraemer et al. 2007). The var gene repertoire is divided into three major groups (A, B, and C) (Gardner et al. 2002; Kraemer & Smith 2003; Lavstsen et al. 2003). However, unlike in *T. congolense*, hybridisation between groups is possible and frequent, being described by the intermediate groups B/A and B/C (Lavstsen et al. 2003). This categorisation, which is based on both 5' upstream sequence similarities

and chromosomal location, was subsequently updated to groups A1-2, B1-B4, C1-2, and E, as a means to include 'atypical' variants generated by the frequent recombination and chromosomal loci movement (Kraemer et al. 2007).

In *P. falciparum*, telomeric clustering is thought to be stochastic, thus allowing *var* genes to recombine with a multitude of other potentially unrelated *var* genes, and exponentially increasing the potential for variant formation (Freitas-Junior et al. 2000). However, particular chromosomal end combinations may be more efficient than others at driving recombination due to higher sequence homology and/or peculiarities of chromatin structure (Freitas-Junior et al. 2000). In fact, Kraemer et al. (2007) observed that recombination preferentially occurs within *var* groups, except amongst the *var* groups B and C of the core chromosomes (central *var* clusters). These two groups are more similar to each other, therefore chimaeric genes are more common within these clades, suggesting that group B and C recombine more often with each other than with *var* group A (Kraemer et al. 2007). These differences in recombination probability observed amongst different *var* groups are referred to as gene recombination hierarchy (Kraemer et al. 2007). Such a concept can be transposed to *T. congolense* as we too see high prevalence of recombination within groups (i.e. phylotypes), followed by scarce recombination between groups of higher-similarity (e.g. phylotypes 8 and 4) (**Figure 64**). The formation of distinct antigen groups in the genome has been considered a response to immune selection pressures (Gupta et al. 1996). If that applies to *T. congolense*, then the presence of inter-phylotype recombination would not be advantageous because it would counteract such process.

With such key differences in repertoire diversity generation, it might be expected that the VSG repertoires of the three species diverge at different rates. In *T. brucei*, the VSG repertoires change rapidly over time, even within the same isolate. For example, one of the genomes used in this study, 427var3, is an uncultured variant of the strain Lister 427, used as reference here. Yet, only 34.62 % of their VSG repertoire is fully coupled, suggesting that over time Lister 427 has drastically changed its VSG repertoire. Furthermore, whilst in *T. congolense* the percentage of fully coupled VSGs between the reference and a field strain can have a wide range (5 % to 69 %), in *T. brucei*, even closely related strains have dissimilar VSG repertoires up to a limit of 12 % identity (12 % to 37 %). This suggests that gene turnover is occurring at a rate fast enough to homogenize the VSG gene pool. These results corroborate the view of Hall et al (2013) that *T. brucei* antigenic variation is

built continuously by many variants, rather than in a tightly regulated, intermittent manner. In *T. vivax*, we observe the opposite situation; with the exception of three isolates, at least 50 % of VSGs are conserved in the full extent of the gene, clearly confirming that ectopic recombination amongst *T. vivax* VSGs is much less frequent than in its relatives.

Recombination in *T. brucei* and *T. congolense* may also be occurring at meiosis. Mating and meiosis as mechanisms of gene exchange have been well characterised in *T. brucei* (Turner et al. 1990; Tait et al. 2007; Peacock et al. 2011; Peacock, Bailey, et al. 2014). In *T. congolense*, mating has been shown amongst *T. congolense* savannah in two distinct occasions (Morrison, Tweedie, et al. 2009; Tihon et al. 2017). *T. vivax*, however, despite having the machinery for sexual recombination and therefore the potential to mate, appears to be clonal (Duffy et al. 2009). Clonality may have further implications for the VSG repertoire. For example, the *T. vivax* VSG repertoire presents a strong geographic signature that directly mirrors population structure (chapter 5). Such strong signal may reflect both the clonal nature of the parasite and the scarcity of ectopic recombination within the VSG repertoire. In African trypanosomes undergoing sexual exchange, VSG repertoire not always replicates the population structure. For instance, in *T. congolense* the VSG repertoire appears uncoupled from the rest of the genome (chapter 2), most likely as a result of a combination of non-Mendelian inheritance of subtelomeres, gene flow and ectopic recombination.

The key aspect is why the genes within these families diverged so much that recombination stopped being possible between lineages. Perhaps it is linked to *T. vivax* losing sexual replication. The presence of meiotic machinery in *T. vivax* suggests it is ancestral in nature, being therefore more likely that clonality reflects a *T. vivax* loss of function rather than a *T. congolense/T. brucei* gain (Duffy et al. 2009). Hypothetically, the ancestor maintained a diverse VSG repertoire, driven by sexual recombination and ectopic gene conversion, which originated a multitude of distinct VSG. As *T. vivax* diverged and became asexual, VSG recombination ceased, and VSG repertoire evolved through changes in gene copy number rather than mosaicism. Conversely, the sexual nature of *T. brucei* and *T. congolense* are likely to have contributed to greater recombination levels, albeit in different ways. After speciation, *T. brucei* seems to have devised strategies to accelerate recombination, possibly for the purpose of infection chronicity and efficient transmission, whilst *T. congolense* devised different fitness strategies potentially involving deceleration of

diversifying selection for the preservation of the distinct lineages. Such an approach would be beneficial for example if these were functional differentiated or developmentally regulated lineages.

6.4.1 Conclusion

This chapter described the differences in VSG diversity generation in the different African trypanosome species from a population perspective. The work presented here shows species-specific signatures of gene hypervariability, represented by clear differences in the average recombination trait of *T. brucei*, *T. congolense* and *T. vivax* of VSG sequences. The *T. vivax* VSG repertoire is for the most part conserved across the species, preserving high sequence orthology even between distantly related isolates. In contrast, the *T. brucei* VSG repertoire is dynamic, highly recombinant, even within isolates of shared backgrounds. *T. congolense* is in between, preserving high levels of recombination, but keeping it restricted to VSGs of the same phylotype, perhaps reflecting its intermediate positioning in the trypanosome phylogeny. In conclusion, I expect the overall levels of VSG orthology among strains of each species to be alike for *T. brucei* and *T. congolense* (40 % and 41 %, respectively), but much higher for *T. vivax* (74 %).

Chapter 7. General Discussion

It is the view in some quarters that HAT will be eradicated in the very near future. The success of continental-wide control programmes enforced in the past 18 years predicts that elimination of HAT by 2020 will be possible (Holmes 2014; Holmes 2015). These programmes were based on detailed surveillance, investment on diagnostics and treatment availability, as well as improved epidemiological mapping of the disease (Holmes 2014). Needless to say, efforts to achieve eradication must continue, to prevent re-emergence of the disease, as observed at the end of the 20th century (Holmes 2014). Nonetheless, this position reflects a considerable effort to understand and combat the human-infecting trypanosomes over the last few decades. The focus on HAT by African governments, the WHO, and research funders has resulted in outstanding models of disease and deep understanding of the *T. brucei gambiense* and *T. brucei rhodesiense* biology, distribution and diversity. In stark contrast, AAT remains poorly measured, coarsely described and largely unstudied in research laboratories. It follows that innovations in prevention and treatment of AAT are rare, and the situation in the field has barely progressed, even while the effort against HAT has increased.

Whilst AAT is often seen as a veterinary extension of HAT and treated as a single disease, the reality is that AAT is a spectrum of diseases, caused by multiple species and strains of African trypanosomes (Morrison et al. 2016). This results in large variability in pathogenesis, epidemiology, and clinical outcome. This is shown by the differences in virulence within *T. congolense* as sub-type ‘savannah’ is more pathogenic than sub-types ‘forest’ and ‘kilifi’ (Gibson 2007; Auty et al. 2015), or by the differences in clinical outcome in *T. vivax* infections. In East Africa, *T. vivax*, which usually causes mild, chronic disease, can sporadically result in acute haemorrhagic syndromes (Welde et al. 1983). In Brazil, whilst most *T. vivax* infections in endemic areas result in chronic disease of low parasitaemia, localised epidemics causing high mortality rates have been associated with particular isolates, such as the ‘Lins’ strain described in chapter 5 and also by Cadioli et al. (2012) and Fidelis Jr et al. (2016). Although anecdotal evidence associating particular disease symptoms and geographical foci with disease pathology is plentiful among local veterinarians and farmers, a real understanding of the association between disease genotypes and phenotypes is still lacking.

As the main surface protein interacting with the host immune system, it would be naïve to think that VSGs have the sole purpose of providing raw material for antigenic variation. In fact, VSGs have been shown to have other roles in pathogenesis, particularly in the modulation of the host's cellular responses (Stijlemans et al. 2016). For instance, VSGs specifically inhibit activation of the alternative pathway in *T. brucei gambiense*, preventing trypanosome lysis (Devine et al. 1986). VSGs also induce the pro-inflammatory response, which causes many of the disease symptoms. For example, as a response to stress, GPI-Phospholipase C (GPI-PLC), an endogenous phospholipase, is activated to cleave the VSG off the GPI anchor, resulting in the release of soluble VSGs (De Almeida & Turner 1983; Bulow & Overath 1986; Ferguson et al. 1988). Both the cleaved GPI anchor and the soluble VSGs trigger the activation of the pro-inflammatory cascade and the expression of pro-inflammatory genes (Leppert et al. 2007). The soluble VSGs also trigger receptors that prime macrophages and induce an IFN- γ response (Magez et al. 1998; Magez et al. 2002; Leppert et al. 2007; Mansfield & Paulnock 2005). This is further exacerbated by increased MHC-II presentation and activation of CD4+ T-cells (Schleifer et al. 1993). The excessive IFN- γ production triggers the acute inflammatory response responsible for acute anaemia development (Stijlemans et al. 2007; Stijlemans et al. 2008; Stijlemans et al. 2010).

Here I have shown multiple times how the three major African trypanosome species differ in terms of their VSG repertoire. Whilst it was traditionally assumed that the VSG machinery is one and the same between African trypanosomes, this work challenges such dogma. Every aspect investigated here [and in other studies (Marcello & Barry 2007a; Jackson et al. 2012; Jackson et al. 2013; Hall et al. 2013; Cross et al. 2014; Jackson et al. 2015; Mugnier et al. 2015)] reveals major differences, be it in repertoire composition, gene expression, sequence diversity or diversity generation. With such diversity among the VSG family, it is at least possible that not all VSGs have the same effect.

Studying diversity may help understanding clinical outcome, but for this, a robust molecular systematics for analysing diversity must exist. The VAP fulfils that need. It provides a platform for researchers to start asking how VSGs affect disease and it lays the basis for systematic studies of genomic and gene expression diversity. It allows researchers to start asking whether the stable differences I observed in *T.*

congolense VSG repertoire translate into phenotypic differences. In the future, the VAP can be used to better understand how VSG determine disease phenotype in a number of ways. For instance, the VAP can be used to identify and characterise programs of VSG expression. Studying patterns of antigenic variation and how VSG expression dynamics change through the course of infection can help explain pathogenesis through the association of particular VSGs or VSG patterns with disease course and syndromes. The existence of VSG expression ‘programs’, based on differential switching probabilities among VSG genes, is a long-standing hypothesis that has never been properly addressed for lack of analytic methods (Barry 1986). Now that even small changes in VSG expression can be detected, quantified and directly compared, this hypothesis can be revisited.

As research improves, the VAP will be empowered by establishing genotype-phenotype associations. Therefore, these data will also have epidemiological and diagnostic value that may be employed to improve AAT disease mapping. If certain VSGs or VSG expression patterns are linked to particular clinical outcomes, they become epidemiological markers. Such information can be used for resource management. For example, in high-risk areas, drugs can be quickly directed to infections predicted to cause acute, lethal disease rather than those linked to mild disease are circulating in the area. Similarly, epidemiological mapping of disease can help targeting vector control strategies to areas where virulent genotypes are circulating.

The characterisation of VSG expression hierarchies can help identify early-stage VSGs. If employed in a multivalent vaccine, a rapid immune response to these VSG before the initial rise of parasitaemia in the bloodstream may be successfully control infection. Better still, the VAP can help identifying constitutively expressed mVSGs. Metacyclic-specific, highly expressed VSGs may be good candidates for vaccination because they would prime the immune system to fight disease before the establishment of a bloodstream infection and the resulting immune suppression. In chapter 3, I have demonstrated how the VAP provides a way to functionally discriminate among VSGs and identify constitutively expressed genes. I discovered that the mVSG expression profile in *T. congolense* isolate Tc148 is non-random, reproducible, and distinct from the genomic profile. I have identified phylotype 8 genes as potentially preferentially expressed in the metacyclic life stage, which agrees with previous reports in the literature characterising mVSG expression in *T. congolense* (Eshita et al. 1992; Helm et al. 2009). Although questions still remain

about their expression profile in natural fly populations, these genes might be good candidates to include in a multivalent vaccine.

Despite its well-described bloodstream stage, *T. brucei* has recently been discovered in other tissues, such as the skin and the adipose tissue (Capewell et al. 2016; Trindade et al. 2016). Tissue tropism may add another level of immune evasion in which antigenic variation may be different. The VAP can be applied to experimental infections of *T. congolense* and *T. vivax*, such as those previously performed to compare gene expression between parasites in the bloodstream and those in the adipose tissue in *T. brucei* (Trindade et al. 2016), to reveal differences in antigenic expression and variation. For instance, understanding how antigenic variation works in these tissue reservoirs may help explain why some trypanosome species, like *T. vivax* and *T. brucei gambiense*, are able to establish long chronic infections with barely undetectable parasitaemia (Malvy & Chappuis 2011; Fidelis Jr et al. 2016). Perhaps the patterns of VSG expression change when the parasites colonise different tissues. These changes in VSG 'program' might be essential for immune evasion in these tissues and ultimately allow long-term survival.

In all these ways, the VAP could be applied to better understand the phenotypic consequences of VSG genetic and expression diversity. Yet, in terms of *T. brucei*, a VSG systematics is still lacking. Despite all the molecular and cell biology advances in *T. brucei*, its extremely dynamic VSG repertoire is a challenge for profiling approaches. Future research attempting to profile *T. brucei* VSG based on amino acid signatures may prove difficult because of the extreme degree of mosaicism (Marcello & Barry 2007b) and the ability to convert genes between very diverse donor regions (Hall et al. 2013). *T. brucei* VSG passed through a bottleneck selection, from which they emerged structurally homogenous and recombination-prone (Jackson & Barry 2012). The *T. brucei* VSG repertoire is highly mosaic, even within isolates of shared backgrounds, as shown by the high levels of unmapped and multi-coupled VSGs shown in chapter 6. This dynamism and low sequence orthology between strains precludes a systematics based on conserved VSG or long VSG motifs. To overcome these challenges, a *T. brucei* VAP methodology may have to resort to the analysis of short amino acid motif combinations in a gene network analysis. Gene network analysis [reviewed in Emmert-Streib et al. (2014)] has had several successful implementations, particularly in immunology (Nacu et al. 2007), and cancer research (Madhamshettiwar et al. 2012). For *T. brucei*, this could potentially be achieved by identifying mosaic networks across the population

because mosaics formed from the same set of donors can have higher nucleotide identity than their donors (Hall et al. 2013). Therefore, the degree of association between gene segments may be more informative than the gene segments themselves.

Application of variant antigen profiling, for all trypanosome species, in the ways described above would likely transform our understanding. Similar knowledge of surface protein diversity has already led to substantial progress in other infectious diseases. In pregnancy-associated malaria, which causes anaemia, premature birth, and increased neonatal mortality, *Plasmodium falciparum* parasites colonise the placental tissues. While several genes are involved (Tuikue Ndam et al. 2008b), one of main processes in this tropism is the adhesion of infected erythrocytes to glycosaminoglycan chondroitin sulphate A (CSA) in the intervillous space of the placenta. This adhesion, which is mediated by *var2csa* (Rowe et al. 2002; Salanti et al. 2003), allows the infected erythrocytes to accumulate in the placenta whilst avoiding the spleen-mediated surveillance mechanisms. The *var2csa* gene is conserved among parasite strains (Salanti et al. 2003) and its transcription is significantly up-regulated in placental isolates compared to bloodstream isolates (Salanti et al. 2003; Tuikue Ndam et al. 2008a). After identification of this specific *var* gene, several studies have been developed with the aim to characterise binding activity, antigenicity and immunogenicity [reviewed in Fried & Duffy (2015)]. This work resulted in two VAR2CSA-based vaccine candidates currently in the path for human clinical studies (Fried & Duffy 2015). However, due to the large variation in clinical outcomes in malarial infections, understanding the diversity of the *var2csa* gene and its relationship to different disease phenotypes has been and will continue to be crucial (Patel et al. 2017).

In influenza, surface antigen profiling has improved the prediction of antigenic drift required for the biannual re-evaluation of the flu vaccine (McHardy & Adams 2009). Influenza A viruses are classified into different sub-types according to the specific genetic combination of their surface glycoproteins [hemagglutinin (H1–16) and neuraminidase (N1–9)]. A vaccine, based on inactivated viruses, is available and elicits an antibody response targeting mainly the hemagglutinin glycoprotein. Due to the high frequency of antigenic drift in Influenza A, the vaccine is of limited protection. This inspired a global surveillance program to detect emerging antigenic variants and predict whether the changes in the surface glycoproteins would require an update of the existing vaccine strain. Such analysis is only possible because

large-scale genomics and antigenic typing identified patterns of viral circulation and antigenic diversity evolution (Smith et al. 2004; Russell et al. 2008).

There are more examples of vaccines targeting surface proteins. In dengue, the envelope protein, a surface protein, is the preferred candidate antigen for the development of recombinant protein, DNA and subunit vaccines (Liu et al. 2016). Recent examples include the E-based recombinant subunit vaccine V180 from Merck, which is currently in phase I clinical trials (Coller et al. 2011), and the tetravalent DNA vaccine based on the envelope genes of multiple Dengue virus serotypes (Raviprakash et al. 2000; Raviprakash et al. 2001; Blair et al. 2006). In *Staphylococcus aureus* infections, the most common cause of hospital acquired infections, vaccine assembly from surface proteins, has been extensively discussed (Stranger-Jones et al. 2006). In particular, one of the approaches was to target a combination of four surface antigens that induce an antibody response and may provide protection (Stranger-Jones et al. 2006). Simultaneously, research into the immunogenicity potential of SpA, a surface antigen constitutively expressed by all clinical isolates, is under way (Missiakas & Schneewind 2016).

All these examples show that understanding the diversity of a pathogen's surface proteins and how they affect virulence can help the development of sustainable solutions for various diseases. The VAP provides the framework for such development in AAT.

Translating the VAP into an epidemiologically-relevant metric will require, as noted, a much better definition of clinical outcome, linked to VSG profile, be that genomic or transcriptomic. Yet it will also require the translation of genomics into the clinical setting. One aspect that distinguishes this thesis is the analysis of natural populations rather than laboratory-adapted isolates. Antigenic variation is a complex phenomenon affected by an evolutionary arms race among the parasite, the vector, and the host, where the importance of species differences and gene diversity cannot be overstated. Working with field isolates has traditionally been confined to analyses of diversity of conserved markers and prevalence studies due to low parasitaemia in natural infections (and the consequent requirement of animal passage) and the great level of unexpected variation encountered. Nonetheless, experimental models of disease may not accurately represent the natural patterns of antigenic variation. Therefore, it is important that the advances in the laboratory-based molecular and cell biology are accompanied by field-based research.

Fortunately, the improvements in sequencing technology in recent years, such as the portable Oxford Nanopore technology and the selective sequencing technologies developed by multiple companies, may help bring research back to natural disease settings. Selective whole-genome amplification has been successfully implemented in *P. falciparum* (Oyola et al. 2016; Rutledge & Ariani 2017), albeit good coverage of the *var* gene repertoire is still to be achieved because the probes designed targeted AT-rich regions, despite the subtelomeres having a higher GC content (Oyola et al. 2016). A more recent approach is target enrichment via hybridization-based capture (e.g. Agilent's SureSelect Target Enrichment System or the myBaits target capture kits from Arbor Biosciences), which has been widely used for exome sequencing in the study of human biology and human diseases [for example in cancer research (Jones et al. 2010), in diabetes (Bonnetfond et al. 2010), and in Mendelian disorders (Ng et al. 2009)]. In the infectious diseases field, it is still underexploited, but has great potential because the use of complex baits to capture selected DNA or RNA allows unprecedented enrichment of microbial genomes or transcriptomes from field samples.

In trypanosomatids, multiplexed splice-leader sequencing has been developed for the transcriptome sequencing of *Leishmania donovani* clinical isolates (Cuypers et al. 2017). This strategy offers an alternative to poly-A tail mRNA purification methods that is specific for parasite RNA, thus resulting in more efficient, deeper sequencing. This approach is ideal for parasite sequencing directly from field isolates. If such method can be applied to African trypanosome genomes, the possibilities for the analysis of VSG diversity in natural populations are endless. With these advances in methodology and technology, trypanosome research can return to the field and focus on genomic epidemiology and how genetic diversity links to disease phenotype, geography and transmission. It can also move towards understanding sylvatic and domestic transmission cycles, how they affect strain and species diversity in different regions, and how they associate to the different disease hosts. In fact, within trypanosome research, *T. cruzi* provides evidence of the benefits of studying disease in natural settings. *T. cruzi* genotyping in the field from both host and vector samples has been important to understand the complex eco-epidemiology of Chagas disease and the relationship between parasite genotypes, geography and disease outcome [reviewed by Messenger et al. (2015) and Zingales (2017)].

I anticipate great developments in the trypanosomiasis field research. The combination of genomics, transcriptomics and tools like the VAP will allow fast and

accurate analysis of complex big data. We can now ask more questions because we have the means to answer them. Personally, I am intrigued by the possibility of antigenic synchronicity among infected cattle of the same herd, or by the patterns of antigenic variation among trypanotolerant cattle. The future will certainly bring the opportunity to explore all these questions and many more, like how antigenic variation continues in chronic, natural infections of wild animals, where parasitaemia is barely detectable. Ultimately, the ability to take research from the lab back to the field is extraordinary. It allows the integration of local knowledge with the cutting-edge methodological advances, it brings back awareness to where the problem is, and it promotes research that is focused on sustainable solutions to control trypanosomiasis and improve animal welfare and economic development.

7.1 Conclusion

In this thesis I have shown how VSGs differ among African trypanosomes in terms of diversity, expression and evolution, and how their differences may be used to develop automated methods of VSG profiling from sequence data. Together with previously described differences in structure and codon bias, it is clear that while there may be a common phenotype of antigenic variation in all species, the mechanisms of antigenic switching and diversification are different. Variability in the antigenic variation system is likely to have had a considerable impact on current VSG expression patterns, transmission, disease progression, and virulence. Adaptation strategies are tightly dependent on interactions with the vector, host, parasite and environment, but they also associate closely with the degree of population structuring, mating frequency and patterns, and potentially VSG hierarchies. The need for further study of all of these aspects in *T. congolense* and *T. vivax* is clear and unquestionable. The selective pressures associated with VSG diversity generation in *T. vivax* and how they may relate to the different clinical outcomes also need addressing. The use of *T. brucei* as a model for antigenic variation has given us a deep understanding of VSG expression and switching in that species. However, the evidence that this might not apply to its closest relatives necessitates studies to understand how antigenic variation can be generated in different ways. Advances in functional transcriptomics and in understanding pathogenesis will be critical in this. Variant antigen profiling will have a role in these experiments, making the composition of VSG repertoires within these studies tractable, and as sequencing technologies evolve in sensitivity and depth, so will the VAP.

References

- Van Den Abbeele, J. et al., 1999. *Trypanosoma brucei* spp. development in the tsetse fly: Characterization of the post-mesocyclic stages in the foregut and proboscis. *Parasitology*, 118(5), pp.469–478.
- Adam, R.D., 2000. The *Giardia lamblia* genome. *International Journal for Parasitology*, 30(4), pp.475–484.
- Adams, E.R. et al., 2010. New *Trypanosoma* (Duttonella) *vivax* genotypes from tsetse flies in East Africa. *Parasitology*, 137(4), pp.641–650.
- Al-Khedery, B. & Allred, D.R., 2006. Antigenic variation in *Babesia bovis* occurs through segmental gene conversion of the ves multigene family, within a bidirectional locus of active transcription. *Molecular Microbiology*, 59(2), pp.402–414.
- Albrecht, L. et al., 2010. The South American *Plasmodium falciparum* var gene repertoire is limited, highly shared and possibly lacks several antigenic types. *Gene*, 453(1–2), pp.37–44.
- Alexandre, S. et al., 1996. Families of adenylate cyclase genes in *Trypanosoma brucei*. *Molecular and Biochemical Parasitology*, 77(2), pp.173–182.
- Alexandre, S. et al., 1988. Putative genes of a variant-specific antigen gene transcription unit in *Trypanosoma brucei*. *Molecular and Cellular Biology*, 8(6), pp.2367–78.
- Allred, D.R. et al., 1990. Molecular basis for surface antigen size polymorphisms and conservation of a neutralization-sensitive epitope in *Anaplasma marginale*. *Proceedings of the National Academy of Sciences of the United States of America*, 87(8), pp.3220–3224.
- De Almeida, M.L.C. & Turner, M.J., 1983. The membrane form of variant surface glycoproteins of *Trypanosoma brucei*. *Nature*, 302(5906), pp.349–352.
- Alsford, S. et al., 2001. Diversity and dynamics of the minichromosomal karyotype in *Trypanosoma brucei*. *Molecular and Biochemical Parasitology*, 113(1), pp.79–88.
- Alsford, S. et al., 2012. Epigenetic mechanisms, nuclear architecture and the control of gene expression in trypanosomes. *Expert Reviews in Molecular Medicine*, 14(May), pp.1–20.
- Altschul, S.F. et al., 1990. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3), pp.403–10.
- Alves-Silva, J. et al., 2010. An insight into the sialome of *Glossina morsitans morsitans*. *BMC Genomics*, 11, p.213.
- Amiguet-Vercher, A. et al., 2004. Loss of the mono-allelic control of the VSG expression sites during the development of *Trypanosoma brucei* in the bloodstream. *Molecular Microbiology*, 51(6), pp.1577–1588.
- Anderson, T.J. et al., 2000. Microsatellite markers reveal a spectrum of population structures in the malaria parasite *Plasmodium falciparum*. *Molecular Biology and Evolution*, 17(10),

- pp.1467–1482.
- Arcà, B. et al., 2007. An insight into the sialome of the adult female mosquito *Aedes albopictus*. *Insect Biochemistry and Molecular Biology*, 37(2), pp.107–127.
- Arenas, M. & Posada, D., 2010. Coalescent simulation of intracodon recombination. *Genetics*, 184(2), pp.429–437.
- Assumpção, T.C.F. et al., 2008. An insight into the sialome of the blood-sucking bug *Triatoma infestans*, a vector of Chagas' disease. *Insect Biochemistry and Molecular Biology*, 38(2), pp.213–232.
- Authié, E., Muteti, D.K. & Williams, D.J., 1993. Antibody responses to invariant antigens of *Trypanosoma congolense* in cattle of differing susceptibility to trypanosomiasis. *Parasite Immunology*, 15(2), pp.101–111.
- Auty, H. et al., 2015. Cattle trypanosomosis: the diversity of trypanosomes and implications for disease epidemiology and control. *Revue Scientifique et Technique*, 34(2), pp.587–598.
- Van der Auwera, G.A. et al., 2013. From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. *Current Protocols in Bioinformatics*, (SUPL.43).
- Ayisheshim, A. et al., 2015. Review on Bovine Trypanosomosis in Ethiopia. *Acta Parasitologica Globalis*, 6(3), pp.136–146.
- Bachmann, A. et al., 2016. Mosquito Passage Dramatically Changes *var* Gene Expression in Controlled Human *Plasmodium falciparum* Infections. *PLoS Pathogens*, 12(4), p.e1005538.
- Balber, A.E. et al., 1979. Inactivation or elimination of potentially trypanolytic, complement-activating immune complexes by pathogenic trypanosomes. *Infection and Immunity*, 24(3), pp.617–627.
- Barbet, A.F. & Kamper, S.M., 1993. The importance of mosaic genes to trypanosome survival. *Parasitology Today*, 9(2), pp.63–66.
- Barnes, R.L. & McCulloch, R., 2007. *Trypanosoma brucei* homologous recombination is dependent on substrate length and homology, though displays a differential dependence on mismatch repair as substrate length decreases. *Nucleic Acids Research*, 35(10), pp.3478–3493.
- Barnwell, E.M. et al., 2010. Developmental regulation and extracellular release of a VSG expression-site-associated gene product from *Trypanosoma brucei* bloodstream forms. *Journal of Cell Science*, 123(19), pp.3401–3411.
- Barrett, M.P. & Fairlamb, A.H., 1999. The Biochemical Basis of Arsenical – Diamidine Crossresistance in African Trypanosomes. *Parasitology Today*, 15(4), pp.136–140.
- Barry, A.E. et al., 2007. Population genomics of the immune evasion (*var*) genes of *Plasmodium falciparum*. *PLoS Pathogens*, 3(3), p.e34.
- Barry, J.D., 1986. Antigenic variation during *Trypanosoma vivax* infections of different host species. *Parasitology*, 92 (Pt 1)(May), pp.51–65.

- Barry, J.D. et al., 1998. VSG gene control and infectivity strategy of metacyclic stage *Trypanosoma brucei*. *Molecular and Biochemical Parasitology*, 91(1), pp.93–105.
- Barry, J.D. et al., 2005. What the genome sequence is revealing about trypanosome antigenic variation. *Biochemical Society transactions*, 33(Pt 5), pp.986–9.
- Barry, J.D. et al., 2003. Why are parasite contingency genes often associated with telomeres? *International Journal for Parasitology*, 33(4), pp.29–45.
- Barry, J.D., Crowe, J.S. & Vickerman, K., 1983. Instability of the *Trypanosoma brucei* rhodesiense metacyclic variable antigen repertoire. *Nature*, 306, pp.699–701.
- Barry, J.D. & McCulloch, R., 2001. Antigenic variation in trypanosomes: enhanced phenotypic variation in a eukaryotic parasite. *Advances in parasitology*, 49, pp.1–70.
- Bartholomeu, D.C. et al., 2009. Genomic organization and expression profile of the mucin-associated surface protein (masp) family of the human pathogen *Trypanosoma cruzi*. *Nucleic Acids Research*, 37(10), pp.3407–3417.
- Baruch, D.I. et al., 1995. Cloning the *P. falciparum* gene encoding PfEMP1, a malarial variant antigen and adherence receptor on the surface of parasitized human erythrocytes. *Cell*, 82(1), pp.77–87.
- Bastos, T.S.A. et al., 2017. First outbreak and subsequent cases of *Trypanosoma vivax* in the state of Goiás, Brazil. *Revista Brasileira de Parasitologia Veterinária*, 2961(2012), pp.366–371.
- Batista, J.S. et al., 2012. Highly debilitating natural *Trypanosoma vivax* infections in Brazilian calves: Epidemiology, pathology, and probable transplacental transmission. *Parasitology Research*, 110(1), pp.73–80.
- Batista, J.S. et al., 2009. Infection by *Trypanosoma vivax* in goats and sheep in the Brazilian semiarid region: From acute disease outbreak to chronic cryptic infection. *Veterinary Parasitology*, 165(1–2), pp.131–135.
- Batista, J.S. et al., 2007. Trypanosomiasis by *Trypanosoma vivax* in cattle in the Brazilian semiarid: Description of an outbreak and lesions in the nervous system. *Veterinary Parasitology*, 143(2), pp.174–181.
- Batram, C. et al., 2014. Expression site attenuation mechanistically links antigenic variation and development in *Trypanosoma brucei*. *eLife*, 3, p.e02324.
- Becker, M. et al., 2004. Isolation of the repertoire of VSG expression site containing telomeres of *Trypanosoma brucei* 427 using transformation-associated recombination in yeast. *Genome Research*, 14(11), pp.2319–2329.
- Bell, J.S. & McCulloch, R., 2003. Mismatch Repair Regulates Homologous Recombination, but Has Little Influence on Antigenic Variation, in *Trypanosoma brucei*. *Journal of Biological Chemistry*, 278(46), pp.45182–45188.
- Berberof, M., Pérez-Morga, D. & Pays, E., 2001. A receptor-like flagellar pocket glycoprotein specific to *Trypanosoma brucei gambiense*. *Molecular and Biochemical Parasitology*, 113(1), pp.127–138.
- Bernards, A. et al., 1981. Activation of trypanosome surface glycoprotein genes involves a

- duplication-transposition leading to an altered 3' end. *Cell*, 27(3) (PT 2), pp.497–505.
- Berriman, M. et al., 2002. The architecture of variant surface glycoprotein gene expression sites in *Trypanosoma brucei*. *Mol Biochem Parasitol*, 122(2), pp.131–140.
- Berriman, M. et al., 2005. The genome of the African trypanosome *Trypanosoma brucei*. *Science*, 309(5733), pp.416–422.
- Birhanu, H. et al., 2015. Surra Sero K-SeT, a new immunochromatographic test for serodiagnosis of *Trypanosoma evansi* infection in domestic animals. *Veterinary Parasitology*, 211(3–4), pp.153–157.
- Black, S.J., Hewett, R.S. & Sendashonga, C.N., 1982. *Trypanosoma brucei* variable surface antigen is released by degenerating parasites but not by actively dividing parasites. *Parasite Immunology*, 4, pp.233–244.
- Blair, P.J. et al., 2006. Evaluation of immunity and protective efficacy of a dengue-3 premembrane and envelope DNA vaccine in *Aotus nancymae* monkeys. *Vaccine*, 24(9), pp.1427–1432.
- Blum, M.L. et al., 1993. A structural motif in the variant surface glycoproteins of *Trypanosoma brucei*. *Nature*, 362, pp.603–609.
- Bonnefond, A. et al., 2010. Molecular diagnosis of neonatal diabetes mellitus using next-generation sequencing of the whole exome. *PLoS ONE*, 5(10), p.e13630.
- Boothroyd, C.E. et al., 2009. A yeast-endonuclease-generated DNA break induces antigenic switching in *Trypanosoma brucei*. *Nature*, 459(7244), pp.278–281.
- Borst, P., 1986. Discontinuous Transcription and Antigenic Variation in Trypanosomes. *Ann. Rev. Biochem.*, 55, pp.701–732.
- van den Bossche, P. et al., 2011. Virulence in *Trypanosoma congolense* Savannah subgroup . A comparison between strains and transmission cycles. *Parasite Immunology*, 33(8), pp.456–460.
- Brayton, K.A. et al., 2002. Antigenic variation of *Anaplasma marginale* msp2 occurs by combinatorial gene conversion. *Molecular Microbiology*, 43(5), pp.1151–1159.
- Bridges, D.J. et al., 2008. Characterisation of the plasma membrane subproteome of bloodstream form *Trypanosoma brucei*. *Proteomics*, 8(1), pp.83–99.
- Bruen, T.C., Philippe, H. & Bryant, D., 2006. A Simple and Robust Statistical Test for Detecting the Presence of Recombination. *Genetics*, 172(April), pp.2665–2681.
- Brun, R., Hecker, H. & Lun, Z.R., 1998. *Trypanosoma evansi* and *T. equiperdum*: Distribution, biology, treatment and phylogenetic relationship (a review). *Veterinary Parasitology*, 79(2), pp.95–107.
- Bull, P.C. et al., 2005. *Plasmodium falciparum* Variant Surface Antigen Expression Patterns during Malaria. *PLoS Pathogens*, 1(3), p.e26.
- Bulow, R. & Overath, P., 1986. Purification and characterization of the membrane-form variant surface glycoprotein hydrolase of *Trypanosoma brucei*. *Journal of Biological Chemistry*, 261(25), pp.11918–11923.
- Buscher, P. et al., 2013. New rapid tests for antibody detection in *Trypanosoma brucei*

- gambiense* sleeping sickness. *N. Engl. J. Med.*, 368, pp.1069–1070.
- Büscher, P. et al., 2014. Sensitivity and specificity of HAT Sero-K-SeT, a rapid diagnostic test for serodiagnosis of sleeping sickness caused by *Trypanosoma brucei gambiense*: A case-control study. *The Lancet Global Health*, 2(6), pp.359–363.
- Cadioli, F.A. et al., 2012. First report of *Trypanosoma vivax* outbreak in dairy cattle in São Paulo state, Brazil. *Revista Brasileira Parasitologia Veterinária*, 21(2), pp.118–124.
- Callejas, S. et al., 2006. Hemizygous subtelomeres of an African trypanosome chromosome may account for over 75% of chromosome length. *Genome Research*, 16(9), pp.1109–1118.
- Campillo, N. & Carrington, M., 2003. The origin of the serum resistance associated (SRA) gene and a model of the structure of the SRA polypeptide from *Trypanosoma brucei rhodesiense*. *Molecular and Biochemical Parasitology*, 127(1), pp.79–84.
- Capbern, A. et al., 1977. *Trypanosoma* au Cours *equiperdum* : Étude de la Trypanosomose des Variations Experimentale Antigéniques du Lapin. *Experimental parasitology*, 42, pp.6–13.
- Capewell, P. et al., 2016. The skin is a significant but overlooked anatomical reservoir for vector-borne African trypanosomes. *eLife*, 5(September 2016), p.e17716.
- Capewell, P. et al., 2013. The TgsGP gene is essential for resistance to human serum in *Trypanosoma brucei gambiense*. *PLoS Pathogens*, 9(10), p.e1003686.
- Carrington, M. et al., 1991. Variant specific glycoprotein of *Trypanosoma brucei* consists of two domains each having an independently conserved pattern of cysteine residues. *Journal of Molecular Biology*, 221(3), pp.823–835.
- Carvalho, A.U. et al., 2008. Ocorrência de *Trypanosoma vivax* no estado de Minas Gerais. *Arquivo Brasileiro de Medicina Veterinária e Zootecnia*, 60(3), pp.769–771.
- Carver, T.J. et al., 2005. ACT: The Artemis comparison tool. *Bioinformatics*, 21(16), pp.3422–3423.
- Chattopadhyay, A. et al., 2005. Structure of the C-terminal domain from *Trypanosoma brucei* variant surface glycoprotein MITat1.2. *Journal of Biological Chemistry*, 280(8), pp.7228–7235.
- Chaves, I. et al., 1999. Control of variant surface glycoprotein gene-expression sites in *Trypanosoma brucei*. *EMBO Journal*, 18(17), pp.4846–4855.
- Chávez, A.S.O. et al., 2012. Expression patterns of *Anaplasma marginale* Msp2 variants change in response to growth in cattle, and tick cells versus mammalian cells. *PLoS ONE*, 7(4), p.e36012.
- Chen, D.S. et al., 2011. A Molecular Epidemiological Study of var Gene Diversity to Characterize the Reservoir of *Plasmodium falciparum* in Humans in Africa A. C. Gruner, ed. *PLoS ONE*, 6(2), p.e16629.
- Chin, C.-S. et al., 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature Methods*, 10(6), pp.563–569.
- Clarkson, M.J., McCabe, W. & Colina, H.S., 1971. Bovine trypanosomiasis in Venezuela.

- Transactions of the Royal Society of Tropical Medicine and Hygiene*, 65(2), pp.257–258.
- Coller, B.A.G. et al., 2011. The development of recombinant subunit envelope-based vaccines to protect against dengue virus induced disease. *Vaccine*, 29(42), pp.7267–7275.
- Connor, R.J., 1992. The diagnosis, treatment and prevention of animal trypanosomiasis under field conditions. FAO. *Panel of Experts on Ecological and Technical Aspects of the Programme for the Control of African Animal Trypanosomiasis and Related Development 24-26 Jun 1991*.
- Cortez, a P. et al., 2006. The taxonomic and phylogenetic relationships of *Trypanosoma vivax* from South America and Africa. *Parasitology*, 133(Pt 2), pp.159–69.
- Cross, G.A.M., 1990. Cellular and genetic aspects of antigenic variation in Trypanosomes. *Annual Reviews of Immunology*, 8, pp.83–110.
- Cross, G.A.M., Wirtz, L.E. & Navarro, M., 1998. Regulation of vsg expression site transcription and switching in *Trypanosoma brucei*. *Molecular and Biochemical Parasitology*, 91(1), pp.77–91.
- Cross, G. a M., Kim, H.S. & Wickstead, B., 2014. Capturing the variant surface glycoprotein repertoire (the VSGnome) of *Trypanosoma brucei* Lister 427. *Molecular and Biochemical Parasitology*, 195(1), pp.59–73.
- Crowe, J.S. et al., 1983. All metacyclic variable antigen types of *Trypanosoma congolense* identified using monoclonal antibodies. *Nature*, 306, pp.389–391.
- Cuypers, B. et al., 2017. Multiplexed Spliced-Leader Sequencing: A high-throughput, selective method for RNA-seq in Trypanosomatids. *Scientific Reports*, 7(1), pp.1–11.
- D'Archivio, S. et al., 2011. Genetic Engineering of *Trypanosoma* (Dutonella) *vivax* and In Vitro Differentiation under Axenic Conditions. *PLoS Neglected Tropical Diseases*, 5(12), p.e1461.
- Dagenais, T.R. et al., 2009. Processing and presentation of variant surface glycoprotein molecules to T cells in African trypanosomiasis. *Journal of immunology*, 183(5), pp.3344–3355.
- Danecek, P. et al., 2011. The variant call format and VCFtools. *Bioinformatics*, 27(15), pp.2156–2158.
- Daniell, H., 2012. The epigenome of *Trypanosoma brucei*: a regulatory interface to an unconventional transcriptional machine. *Biochimica et Biophysica Acta*, 76(October 2009), pp.211–220.
- Das, S. et al., 2010. Transcriptomic and functional analysis of the *Anopheles gambiae* salivary gland in relation to blood feeding. *BMC Genomics*, 11(1), p.566.
- Dayo, G.K. et al., 2010. Prevalence and incidence of bovine trypanosomosis in an agro-pastoral area of southwestern Burkina Faso. *Research in Veterinary Science*, 88(3), pp.470–477.
- Deitsch, K.W., Lukehart, S.A. & Stringer, J.R., 2009. Common strategies for antigenic variation by bacterial, fungal and protozoan pathogens. *Nature Reviews Microbiology*,

- 7(7), pp.493–503.
- Delespaulx, V. et al., 2008. Molecular tools for the rapid detection of drug resistance in animal trypanosomes. *Trends in Parasitology*, 24(5), pp.236–242.
- Denninger, V. et al., 2010. The FACT subunit TbSpt16 is involved in cell cycle specific control of VSG expression sites in *Trypanosoma brucei*. *Molecular Microbiology*, 78(2), pp.459–474.
- Desquesnes, M. et al., 2001. Detection and identification of *Trypanosoma* of African livestock through a single PCR based on internal transcribed spacer 1 of rDNA. *International Journal for Parasitology*, 31(November), pp.610–614.
- Desquesnes, M. et al., 2013. *Trypanosoma evansi* and surra: A review and perspectives on transmission, epidemiology and control, impact, and zoonotic aspects. *BioMed Research International*, 2013.
- Desquesnes, M. & Dia, M.L., 2004. Mechanical transmission of *Trypanosoma congolense* in cattle by the African tabanid *Atylotus agrestis*. *Experimental Parasitology*, 105(3–4), pp.226–231.
- Desquesnes, M. & Dia, M.L., 2003. *Trypanosoma vivax*: Mechanical transmission in cattle by one of the most common African tabanids, *Atylotus agrestis*. *Experimental Parasitology*, 103(1–2), pp.35–43.
- Devine, D. V, Falk, R.J. & Balber, a E., 1986. Restriction of the alternative pathway of human complement by intact *Trypanosoma brucei* subsp. *gambiense*. *Infection and Immunity*, 52(1), pp.223–229.
- Dickin, S.K. & Gibson, W.C., 1989. Hybridisation with a repetitive DNA probe reveals the presence of small chromosomes in *Trypanosoma vivax*. *Molecular and Biochemical Parasitology*, 33(2), pp.135–142.
- Dirie, M.F. et al., 1993. Comparative studies of *Trypanosoma* (Duttonella) *vivax* isolates from Colombia. *Parasitology*, 106, pp.21–29.
- Dobin, A. et al., 2013. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1), pp.15–21.
- Donelson, J.E., 2003. Antigenic variation and the African trypanosome genome. *Acta tropica*, 85(3), pp.391–404.
- Doyle, J.J. et al., 1980. Antigenic variation in clones of animal-infective *Trypanosoma brucei* derived and maintained in vitro. *Parasitology*, 80(2), pp.359–369.
- DuBois, K.N. et al., 2012. NUP-1 Is a large coiled-coil nucleoskeletal protein in trypanosomes with lamin-like functions. *PLoS Biology*, 10(3), p.e1001287.
- Duffy, C.W. et al., 2013. Population Genetics of *Trypanosoma brucei rhodesiense*: Clonality and Diversity within and between Foci. *PLoS Neglected Tropical Diseases*, 7(11), p.e2526.
- Duffy, C.W. et al., 2009. *Trypanosoma vivax* displays a clonal population structure. *International Journal for Parasitology*, 39(13), pp.1475–1483.
- Duraisingh, M.T. & Horn, D., 2016. Epigenetic Regulation of Virulence Gene Expression in

- Parasitic Protozoa. *Cell Host and Microbe*, 19(5), pp.629–640.
- Echodu, R. et al., 2015. Genetic Diversity and Population Structure of *Trypanosoma brucei* in Uganda: Implications for the Epidemiology of Sleeping Sickness and Nagana. *PLoS Neglected Tropical Diseases*, 9(2), p.e0003353.
- Eddy, S.R., 1998. Profile hidden Markov models. *Bioinformatics*, 14(9), pp.755–763.
- Eddy, S.R., 2004. What is a hidden Markov model? *Nature biotechnology*, 22(10), pp.1315–1316.
- Emmert-Streib, F., Dehmer, M. & Haibe-Kains, B., 2014. Gene regulatory networks and their applications: understanding biological and medical problems in terms of networks. *Frontiers in Cell and Developmental Biology*, 2, p.38.
- Engstler, M. et al., 2007. Hydrodynamic Flow-Mediated Protein Sorting on the Cell Surface of Trypanosomes. *Cell*, 131(3), pp.505–515.
- Eshita, Y. et al., 1992. Metacyclic form-specific variable surface glycoprotein-encoding genes of *Trypanosoma* (Nannomonas) *congolense*. *Gene*, 113(2), pp.139–148.
- Esser, K.M. & Schoenbechler, M.J., 1985. Expression of two variant surface glycoproteins on individual African trypanosomes during antigen switching. *Science*, 229(4709), pp.190–3.
- Esser, K.M., Schoenbechler, M.J. & Gingrich, J.B., 1982. *Trypanosoma rhodesiense* blood forms express all antigen specificities relevant to protection against metacyclic (insect form) challenge. *The Journal of Immunology*, 129, pp.1715–1718.
- Fávero, J.F. et al., 2016. *Trypanosoma vivax* infection in goat in west of Santa Catarina state, Brazil. *Comparative Clinical Pathology*, 25(2), pp.497–499.
- Felsenstein, J., 1989. PHYLIP - Phylogeny inference package - v3.2. *Cladistics*, pp.164–166.
- Ferguson, M. et al., 1988. Glycosyl-phosphatidylinositol moiety that anchors *Trypanosoma brucei* variant surface glycoprotein to the membrane. *Science*, 239(4841), pp.753–759.
- Ferguson, M.A.J. et al., 1986. Biosynthesis of *Trypanosoma brucei* Variant Surface Glycoproteins. *Journal of Biological Chemistry*, 261(1), pp.356–362.
- Ferrante, A. & Allison, A.C., 1983. Alternative pathway activation of complement by African trypanosomes lacking a glycoprotein coat. *Parasite Immunology*, 5(5), pp.491–498.
- Fidelis Jr, O.L. et al., 2016. Evaluation of clinical signs, parasitemia, hematologic and biochemical changes in cattle experimentally infected with *Trypanosoma vivax*. *Brazilian Journal of Veterinary Parasitology*, 2961(1), pp.69–81.
- Figueiredo, L.M. & Cross, G.A.M., 2010. Nucleosomes are depleted at the VSG expression site transcribed by RNA polymerase I in African trypanosomes. *Eukaryotic Cell*, 9(1), pp.148–154.
- Figueiredo, L.M., Cross, G.A.M. & Janzen, C.J., 2009. Epigenetic regulation in African trypanosomes: a new kid on the block. *Nature reviews. Microbiology*, 7(7), pp.504–13.
- Figueiredo, L.M., Janzen, C.J. & Cross, G.A., 2008. A Histone Methyltransferase Modulates Antigenic Variation in African Trypanosomes K. Gull, ed. *PLoS Biology*, 6(7), p.e161.
- Finn, R.D., Clements, J. & Eddy, S.R., 2011. HMMER web server: Interactive sequence

- similarity searching. *Nucleic Acids Research*, 39(Web Server issue): W29–W37.
- Franco, J.R. et al., 2014. Epidemiology of human African trypanosomiasis. *Clinical Epidemiology*, 6(1), pp.257–275.
- Franco, J.R. et al., 2017. Monitoring the elimination of human African trypanosomiasis: Update to 2014. *PLoS Neglected Tropical Diseases*, 11(5), p.e0005585.
- Freitas-Junior, L.H. et al., 2000. Frequent ectopic recombination of virulence factor genes in telomeric chromosome clusters of *P. falciparum*. *Nature*, 407(6807), pp.1018–1022.
- Fried, M. & Duffy, P.E., 2015. Designing a VAR2CSA-based vaccine to prevent placental malaria. *Vaccine*, 33(52), pp.7483–7488.
- Frölich, J., 2017. Genotyping of *Anaplasma phagocytophilum* in natural endemic cycles. Dissertation. LMU München: Faculty of Veterinary Medicine.
- Fussenegger, M. et al., 1997. Transformation competence and type-4 pilus biogenesis in *Neisseria gonorrhoeae* - A review. *Gene*, 192(1), pp.125–134.
- Gadelha, C. et al., 2015. Architecture of a host-parasite interface: complex targeting mechanisms revealed through proteomics. *Molecular & cellular proteomics: MCP*, pp.1911–1926.
- Galiza, G.J.N. et al., 2011. High mortality and lesions of the central nervous system in Trypanosomosis by *Trypanosoma vivax* in Brazilian hair sheep. *Veterinary Parasitology*, 182(2–4), pp.359–363.
- Garcia, H. et al., 2005. The detection and PCR-based characterization of the parasites causing trypanosomiasis in water-buffalo herds in Venezuela. *Annals of Tropical Medicine & Parasitology*, 99(4), pp.359–370.
- Gardner, M.J. et al., 2002. Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature*, 419(6906), pp.498–511.
- Gibbs, C.P. & Cross, G.A., 1988. Cloning and transcriptional analysis of a variant surface glycoprotein gene expression site in *Trypanosoma brucei*. *Mol Biochem Parasitol*, 28(3), pp.197–206.
- Gibson, W., 2001. Molecular characterization of field isolates of human pathogenic trypanosomes. *Tropical Medicine & International Health*, 6(5), pp.401–6.
- Gibson, W., 2007. Resolution of the species problem in African trypanosomes. *International Journal for Parasitology*, 37(8–9), pp.829–838.
- Gibson, W., 2012. The origins of the trypanosome genome strains *Trypanosoma brucei brucei* TREU 927, *T. b. gambiense* DAL 972, *T. vivax* Y486 and *T. congolense* IL3000. *Parasites & Vectors*, 5(1), p.71.
- Gibson, W., Peacock, L. & Hutchinson, R., 2017. Microarchitecture of the tsetse fly proboscis. *Parasites & Vectors*, 10(1), p.430.
- Gibson, W.C., Dukes, P. & Gashumba, J.K., 1988. Species-specific DNA probes for the identification of African trypanosomes in tsetse flies. *Parasitology*, 97(1), pp.63–73.
- Ginger, M.L. et al., 2002. Ex vivo and in vitro identification of a consensus promoter for VSG genes expressed by metacyclic-stage trypanosomes in the tsetse fly. *Eukaryotic cell*,

- 1(6), pp.1000–9.
- Giordani, F. et al., 2016. The animal trypanosomiasis and their chemotherapy: A review. *Parasitology*, 143(14), pp.1862–1889.
- Gjini, E. et al., 2010. Critical Interplay between Parasite Differentiation, Host Immunity, and Antigenic Variation in Trypanosome Infections. *The American Naturalist*, 176(4), pp.424–439.
- Glover, L., Alsford, S. & Horn, D., 2013. DNA break site at fragile subtelomeres determines probability and mechanism of antigenic variation in African trypanosomes. *PLoS Pathogens*, 9(3), p.e1003260.
- Glover, L., Jun, J. & Horn, D., 2011. Microhomology-mediated deletion and gene conversion in African trypanosomes. *Nucleic Acids Research*, 39(4), pp.1372–1380.
- Gómez-Rodríguez, J. et al., 2009. Identification of a parasitic immunomodulatory protein triggering the development of suppressive M1 macrophages during African trypanosomiasis. *The Journal of Infectious Diseases*, 200(12), pp.1849–60.
- Gonzalez, A. et al., 1984. Minichromosomal repetitive DNA in *Trypanosoma cruzi*: its use in a high-sensitivity parasite detection assay. *Proceedings of the National Academy of Sciences of the United States of America*, 81(June), pp.3356–3360.
- Graham, S. V et al., 1990. Distinct, developmental stage-specific activation mechanisms of trypanosome VSG genes. *Parasitology*, 101 Pt 3(1990), pp.361–7.
- Graham, S. V. & Barry, J.D., 1991. Expression site-associated genes transcribed independently of variant surface glycoprotein genes in *Trypanosoma brucei*. *Molecular and Biochemical Parasitology*, 47(1), pp.31–41.
- Graham, S. V., Wymer, B. & Barry, J.D., 1998. A trypanosome metacyclic VSG gene promoter with two functionally distinct, life cycle stage-specific activities. *Nucleic Acids Research*, 26(8), pp.1985–1990.
- Graham, S. V & Barry, J.D., 1995. Transcriptional regulation of metacyclic variant surface glycoprotein gene expression during the life cycle of *Trypanosoma brucei*. *Molecular and Cellular Biology*, 15(11), pp.5945–56.
- Gray, A.R., 1965. Antigenic variation in a strain of *Trypanosoma brucei* transmitted by *Glossina morsitans* and *G. palpalis*. *Journal of General Microbiology*, 41(1965), pp.195–214.
- De Greef, C. & Hamers, R., 1994. The serum resistance-associated (SRA) gene of *Trypanosoma brucei rhodesiense* encodes a variant surface glycoprotein-like protein. *Molecular and Biochemical Parasitology*, 68(2), pp.277–284.
- Greif, G. et al., 2013. Transcriptome analysis of the bloodstream stage from the parasite *Trypanosoma vivax*. *BMC genomics*, 14, p.149.
- Guindon, S. et al., 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Systematic Biology*, 59(3), pp.307–321.
- Guindon, S. & Gascuel, O., 2003. A Simple, Fast, and Accurate Method to Estimate Large

- Phylogenies by Maximum Likelihood. *Systematic Biology*, 52(5), pp.696–704.
- Günzl, A. et al., 2003. RNA Polymerase I Transcribes Procyclin Genes and Variant Surface Glycoprotein Gene Expression Sites in *Trypanosoma brucei*. *Eukaryotic Cell*, 2(3), pp.542–551.
- Gupta, S. et al., 1996. The maintenance of strain structure in populations of recombining infectious agents. *Nature Medicine*, 2(4), pp.437–442.
- Haas, B.J. et al., 2013. De novo transcript sequence reconstruction from RNA-Seq: reference generation and analysis with Trinity. *Nature Protocols*, 8(8), pp.1–43.
- HaileMeskel, T.M., 2016. Trypanosomiasis costs 37 African countries USD 4.5 Billion yearly. *FAO, Food and Agriculture Organization of the United Nations*. Available at: <http://www.fao.org/africa/news/detail-news/en/c/461166/> [Accessed May 25, 2017].
- Haines, L.R., 2013. Examining the tsetse teneral phenomenon and permissiveness to trypanosome infection. *Frontiers in Cellular and Infection Microbiology*, 3(November), p.84.
- Hajduk, S.L. et al., 1981. Antigenic variation in cyclically transmitted *Trypanosoma brucei*. Variable antigen type composition of metacyclic trypanosome populations from the salivary glands of *Glossina morsitans*. *Parasitology*, 83(3), pp.595–607.
- Hall, J.P.J., Wang, H. & Barry, J.D., 2013. Mosaic VSGs and the Scale of *Trypanosoma brucei* Antigenic Variation. *PLoS Pathogens*, 9(7), p.e1003502.
- Haroun, E.M. et al., 2003. A preliminary comparative study on the efficacy of quinapyramine sulphate/chloride and melarsoprol in rats, experimentally infected with *Trypanosoma evansi*. *Bulgarian Journal of Veterinary Medicine*, 6(4), pp.215–221.
- Helaine, S. et al., 2007. 3D structure/function analysis of PilX reveals how minor pilins can modulate the virulence properties of type IV pili. *Proceedings of the National Academy of Sciences*, 104(40), pp.15888–15893.
- Helm, J.R. et al., 2009. Analysis of expressed sequence tags from the four main developmental stages of *Trypanosoma congolense*. *Molecular and Biochemical Parasitology*, 168(1), pp.34–42.
- Henson, J., Tischler, G. & Ning, Z., 2012. Next-generation sequencing and large genome assemblies. *Pharmacogenomics*, 13(8), pp.901–915.
- Hertz-Fowler, C. et al., 2008. Telomeric Expression Sites Are Highly Conserved in *Trypanosoma brucei*. *PLoS ONE*, 3(10), p.e3527.
- Hertz-Fowler, C., Renauld, H. & Berriman, M., 2007. The genome of *Trypanosoma brucei*. In J. D. Barry et al., eds. *Trypanosomes: after the genome*. Horizon Bioscience, Wymondham, United Kingdom., pp. 5–48.
- Hoare, C.A., 1972. *The trypanosomes of mammals: a zoological monograph*. Blackwell Scientific Publications, Oxford, United Kindom.
- Hoek, M. & Cross, G.A.M., 2001. Expression-site-associated-gene-8 (ESAG8) is not required for regulation of the VSG expression site in *Trypanosoma brucei*. *Molecular and Biochemical Parasitology*, 117(2), pp.211–215.

- Hoek, M., Engstler, M. & Cross, G.A.M., 2000. Expression-site-associated gene 8 (ESAG8) of *Trypanosoma brucei* is apparently essential and accumulates in the nucleolus. *Journal of Cell Science*, 113, pp.3959–3968.
- Hoek, M., Zanders, T. & Cross, G.A.M., 2002. *Trypanosoma brucei* expression-site-associated-gene-8 protein interacts with a Pumilio family protein. *Molecular and Biochemical Parasitology*, 120(2), pp.269–283.
- Holetz, F.B. et al., 2010. Protein and mRNA content of TcDHH1-containing mRNPs in *Trypanosoma cruzi*. *The FEBS Journal*, 277(16), pp.3415–3426.
- Holmes, P., 2014. First WHO meeting of stakeholders on elimination of gambiense Human African Trypanosomiasis. *PLoS Neglected Tropical Diseases*, 8(10), p.e3244.
- Holmes, P., 2015. On the road to elimination of Rhodesiense human African trypanosomiasis: first WHO meeting of stakeholders. *PLoS Neglected Tropical Diseases*, 9(4), p.e0003571.
- Hope, M. et al., 1999. Analysis of ploidy (in megabase chromosomes) in *Trypanosoma brucei* after genetic exchange. *Molecular and Biochemical Parasitology*, 104(1), pp.1–9.
- Horn, D., 2014. Antigenic variation in African trypanosomes. *Molecular and Biochemical Parasitology*, 195(2), pp.123–129.
- Horn, D. & Barry, J.D., 2005. The central roles of telomeres and subtelomeres in antigenic variation in African trypanosomes. *Chromosome Research*, 13(5), pp.525–533.
- Hornby, H.E., 1921. Trypanosomes and Trypanosomiasis of Cattle. *Journal of Comparative Pathology and Therapeutics*, 34, pp.211–240.
- Hovel-Miner, G. et al., 2016. A Conserved DNA Repeat Promotes Selection of a Diverse Repertoire of *Trypanosoma brucei* Surface Antigens from the Genomic Archive. *PLoS Genetics*, 12(5), p.e1005994.
- Hudson, J.R., 1944. Acute and Subacute trypanosomiasis in cattle caused by *T. vivax*. *Journal of Comparative Pathology*, 44, pp.108–119.
- Huelsenbeck, J.P. & Ronquist, F., 2001. MRBAYES: Bayesian inference of phylogeny. *Bioinformatics*, 17, pp.754–755.
- Hughes, K. et al., 2007. A novel ISWI is involved in VSG expression site downregulation in African trypanosomes. *The EMBO Journal*, 26(9), pp.2400–2410.
- Hutchinson, O.C. et al., 2007. Variant Surface Glycoprotein gene repertoires in *Trypanosoma brucei* have diverged to become strain-specific. *BMC genomics*, 8, p.234.
- Hutchinson, O.C. et al., 2003. VSG structure: Similar N-terminal domains can form functional VSGs with different types of C-terminal domain. *Molecular and Biochemical Parasitology*, 130(2), pp.127–131.
- Hutchinson, R. & Gibson, W., 2015. Rediscovery of *Trypanosoma* (Pycnomonas) *suis*, a tsetse-transmitted trypanosome closely related to *T. brucei*. *Infection, Genetics and Evolution*, 36, pp.381–388.
- International Glossina Genome Initiative, 2014. Genome sequence of the tsetse fly (*Glossina morsitans*): vector of African trypanosomiasis. *Science*, 344(6182), pp.380–386.

- ITM, 2016. Anatomy *Glossina* (tsetse fly), vector of African sleeping disease or trypanosomiasis. *Illustrated Lecture Notes on Tropical Medicine*.
- Jackson, A.P. et al., 2012. Antigenic diversity is generated by distinct evolutionary mechanisms in African trypanosome species. *Proceedings of the National Academy of Sciences of the United States of America*, 109(9), pp.3416–21.
- Jackson, A.P. et al., 2013. A cell-surface phylome for African trypanosomes. *PLoS Neglected Tropical Diseases*, 7 (3), p.e2121.
- Jackson, A.P. et al., 2015. Global Gene Expression Profiling through the Complete Life Cycle of *Trypanosoma vivax*. *PLoS Neglected Tropical Diseases*, 9(8), p.e0003975.
- Jackson, A.P. et al., 2010. The Genome Sequence of *Trypanosoma brucei gambiense*, Causative Agent of Chronic Human African Trypanosomiasis. *PLoS Neglected Tropical Diseases*, 4(4), p.e658.
- Jackson, A.P. & Barry, J.D., 2012. The Evolution of Antigenic Variation in African Trypanosomes. In L. D. Sibley, B. J. Howlett, & J. Heitman, eds. *Evolution of Virulence in Eukaryotic Microbes*. Wiley-Blackwell, pp. 324–337.
- Jackson, D.G., Windle, H.J. & Voorheis, H.P., 1993. The identification, purification, and characterization of two invariant surface glycoproteins located beneath the surface coat barrier of bloodstream forms of *Trypanosoma brucei*. *Journal of Biological Chemistry*, 268(11), pp.8085–8095.
- Jayawardena, A.N., Waksman, B.H. & Eardley, D.D., 1978. Activation of distinct helper and suppressor T cells in experimental trypanosomiasis. *Journal of Immunology*, 121(2), pp.622–8.
- Jiang, L. et al., 2013. PfSETvs methylation of histone H3K36 represses virulence genes in *Plasmodium falciparum*. *Nature*, 499(7457), pp.223–227.
- Johnson, P.J., Kooter, J.M. & Borst, P., 1987. Inactivation of transcription by UV irradiation of *T. brucei* provides evidence for a multicistronic transcription unit including a VSG gene. *Cell*, 51(2), pp.273–281.
- Jones, S. et al., 2010. Frequent mutations of chromatin remodeling gene ARID1A in ovarian clear cell carcinoma. *Science*, 330(6001), pp.228–231.
- Kamper, S.M. & Barbet, A.F., 1992. Surface epitope variation via mosaic gene formation is potential key to long-term survival of *Trypanosoma brucei*. *Molecular and Biochemical Parasitology*, 53(1–2), pp.33–44.
- Kerry, L.E. et al., 2017. Selective inhibition of RNA polymerase I transcription as a potential approach to treat African trypanosomiasis. *PLoS Neglected Tropical Diseases*, 11(3), p.e0005432.
- Kirchgatter, K. & Del Portillo, H.A., 2002. Association of Severe Noncerebral *Plasmodium falciparum* Malaria in Brazil With Expressed PfEMP1 DBL1 α Sequences Lacking Cysteine Residues. *Molecular Medicine*, 8(1), pp.16–23.
- Kolev, N.G., Günzl, A. & Tschudi, C., 2017. Metacyclic VSG expression site promoters are recognized by the same general transcription factor that is required for RNA polymerase

- I transcription of bloodstream expression sites. *Molecular and Biochemical Parasitology*, 216(July), pp.52–55.
- Kooter, J.M. et al., 1987. The anatomy and transcription of a telomeric expression site for variant-specific surface antigens in *T. brucei*. *Cell*, 51(2), pp.261–272.
- Kosakovsky Pond, S.L. et al., 2011. A random effects branch-site model for detecting episodic diversifying selection. *Molecular Biology and Evolution*, 28(11), pp.3033–3043.
- Kosakovsky Pond, S.L. et al., 2006. GARD: A genetic algorithm for recombination detection. *Bioinformatics*, 22(24), pp.3096–3098.
- Kraemer, S.M. et al., 2007. Patterns of gene recombination shape var gene repertoires in *Plasmodium falciparum*: Comparisons of geographically diverse isolates. *BMC Genomics*, 8, pp.1–18.
- Kraemer, S.M. & Smith, J.D., 2003. Evidence for the importance of genetic structuring to the structural and functional specialization of the *Plasmodium falciparum* var gene family. *Molecular Microbiology*, 50(5), pp.1527–1538.
- Kramer, S. et al., 2010. The RNA helicase DHH1 is central to the correct expression of many developmentally regulated mRNAs in trypanosomes. *Journal of Cell Science*, 123(5), pp.699–711.
- Kroubi, M., Karembe, H. & Betbeder, D., 2011. Drug delivery systems in the treatment of African trypanosomiasis infections. *Expert Opinion Drug Delivery*, 8(6), pp.735–747.
- Kukla, B.A. et al., 1987. Use of species-specific DNA probes for detection and identification of trypanosome infection in tsetse flies. *Parasitology*, 95(1), p.1.
- Kumar, S., Stecher, G. & Tamura, K., 2016. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. *Molecular Biology and Evolution*, p.msw054.
- Kyes, S.A., Kraemer, S.M. & Smith, J.D., 2007. Antigenic variation in *Plasmodium falciparum*: Gene organization and regulation of the var multigene family. *Eukaryotic Cell*, 6(9), pp.1511–1520.
- Kyriacou, H.M. et al., 2006. Differential var gene transcription in *Plasmodium falciparum* isolates from patients with cerebral malaria compared to hyperparasitaemia. *Molecular and Biochemical Parasitology*, 150(2), pp.211–218.
- de la Fuente, J. et al., 2005. Genetic diversity of anaplasma species major surface proteins and implications for anaplasmosis serodiagnosis and vaccine development. *Animal health research reviews / Conference of Research Workers in Animal Diseases*, 6(1), pp.75–89.
- De la Fuente, J. et al., 2003. Characterization of the functional domain of major surface protein 1a involved in adhesion of the rickettsia *Anaplasma marginale* to host cells. *Veterinary Microbiology*, 91(2–3), pp.265–283.
- de Lange, T. et al., 1983. Telomere conversion in trypanosomes. *Nucleic Acids Research*, 11(23), pp.8149–8165.
- Langmead, B. & Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4), pp.357–359.

- Lanham, S.M. & Godfrey, D.G., 1970. Isolation of salivarian trypanosomes from man and other mammals using DEAE-cellulose. *Experimental Parasitology*, 28(3), pp.521–534.
- Larionov, V. et al., 1996. Specific cloning of human DNA as yeast artificial chromosomes by transformation-associated recombination. *Proceedings of the National Academy of Sciences of the United States of America*, 93(1), pp.491–496.
- Larkin, M.A. et al., 2007. Clustal W and Clustal X version 2.0. *Bioinformatics*, 23(21), pp.2947–2948.
- Lavstsen, T. et al., 2003. Sub-grouping of *Plasmodium falciparum* 3D7 var genes based on sequence analysis of coding and non-coding regions. *Malaria journal*, 2(27), pp.1–14.
- Lefort, V., Longueville, J.-E. & Gascuel, O., 2017. SMS: Smart Model Selection in PhyML. *Molecular Biology and Evolution*, pp.4–6.
- Lenardo, M.J. et al., 1986. Metacyclic variant surface glycoprotein genes of *Trypanosoma brucei* subsp. *rhodesiense* are activated in situ, and their expression is transcriptionally regulated. *Molecular and Cellular Biology*, 6(6), pp.1991–1997.
- Leppert, B.J., Mansfield, J.M. & Paulnock, D.M., 2007. The soluble variant surface glycoprotein of African trypanosomes is recognized by a macrophage scavenger receptor and induces I kappa B alpha degradation independently of TRAF6-mediated TLR signaling. *Journal of Immunology*, 179(1), pp.548–56.
- Levine, N.D., 1973. The hemoflagellates. *ND Levine, Protozoan parasites of domestic animals and of man, 2nd ed., Burgess Publishing, Minneapolis*, pp.36–78.
- Li, H., 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv*, 0(0), p.3.
- Li, H. et al., 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16), pp.2078–2079.
- Liu, A.Y.C. et al., 1983. The Transposition Unit of Variant Surface Glycoprotein Gene 118 of *Trypanosoma brucei* - Presence of Repeated Elements at its Border and Absence of Promoter-associated Sequences. *Journal of Molecular Biology*, 167(1), pp.57–75.
- Liu, A.Y.C. et al., 1985. Trypanosome Variant Surface Glycoprotein Genes Expressed Early in Infection. *Journal of Molecular Biology*, 175, pp.383–396.
- Liu, Y., Liu, J. & Cheng, G., 2016. Vaccines and immunization strategies for dengue prevention. *Emerging Microbes & Infections*, 5(7), p.e77.
- Lopes, A.H., 2010. Trypanosomatids: Odd Organisms, Devastating Diseases. *The Open Parasitology Journal*, 4(1), pp.30–59.
- López-Farfán, D. et al., 2014. SUMOylation by the E3 Ligase TbSIZ1/PIAS1 Positively Regulates VSG Expression in *Trypanosoma brucei* K. L. Hill, ed. *PLoS Pathogens*, 10(12), p.e1004545.
- Losos, G.J. & Ikede, B.O., 1972. Review of Pathology of Diseases in Domestic and Laboratory Animals Caused by *Trypanosoma congolense*, *T. vivax*, *T. brucei*, *T. rhodesiense* and *T. gambiense*. *Veterinary Pathology*, 9(1_suppl), pp.1–79.
- Luckins, A., 1992. *Methods for diagnosis of trypanosomiasis in livestock*. FAO. Available at:

- <http://www.fao.org/ag/aga/agap/frg/feedback/war/u6600b/u6600b0a.htm>.
- Luckins, A.G. et al., 1994. Early stages of infection with *Trypanosoma congolense*: Parasite kinetics and expression of metacyclic variable antigen types. *Acta Tropica*, 58(3–4), pp.199–206.
- Luckins, A.G., Rae, P. & Gray, M.A., 1981. Development of local skin reactions in rabbits infected with metacyclic forms of *Trypanosoma congolense* cultured in vitro. *Annals of Tropical Medicine and Parasitology*, 75, pp.563–564.
- Luo, R. et al., 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience*, 1(1), p.18.
- Ma, B. et al., 2003. PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Communications in Mass Spectrometry*, 17(20), pp.2337–2342.
- MacGregor, P. et al., 2011. Transmission stages dominate trypanosome within-host dynamics during chronic infections. *Cell Host and Microbe*, 9(4), pp.310–318.
- Macgregor, P. & Matthews, K.R., 2010. New discoveries in the transmission biology of sleeping sickness parasites: applying the basics. *Journal of Molecular Medicine*, 88, pp.865–871.
- Maclean, L. et al., 2007. Spatially and Genetically Distinct African Trypanosome Virulence Variants Defined By Host- Interferon-Gamma Response. *Journal of Infectious Diseases*, 196(11), pp.1620–1628.
- MacLeod, A. et al., 2000. Minisatellite marker analysis of *Trypanosoma brucei*: Reconciliation of clonal, panmictic, and epidemic population genetic structures. *Proceedings of the National Academy of Sciences of the United States of America*, 97(24), pp.13442–13447.
- Madhamshettiwar, P.B. et al., 2012. Gene regulatory network inference: Evaluation and application to ovarian cancer allows the prioritization of drug targets. *Genome Medicine*, 4(5).
- Magez, S. et al., 1998. The glycosyl-inositol-phosphate and dimyristoylglycerol moieties of the glycosylphosphatidylinositol anchor of the trypanosome variant-specific surface glycoprotein are distinct macrophage-activating factors. *Journal of Immunology*, 160(4), pp.1949–56.
- Magez, S. et al., 2002. VSG-GPI anchors of African trypanosomes: Their role in macrophage activation and induction of infection-associated immunopathology. *Microbes and Infection*, 4(9), pp.999–1006.
- Magona, J.W., Walubengo, J. & Odimin, J.T., 2008. Acute haemorrhagic syndrome of bovine trypanosomosis in Uganda. *Acta Tropica*, 107, pp.186–191.
- Maier, B. et al., 2002. Single pilus motor forces exceed 100 pN. *Proceedings of the National Academy of Sciences*, 99(25), pp.16012–16017.
- Majiwa, P.A.O. et al., 1986. Minichromosomal variable surface glycoprotein genes and molecular karyotypes of *Trypanosoma* (Nannomonas) *congolense*. *Gene*, 41(2–3),

pp.183–192.

- Majiwa, P.A.O. et al., 1982. Two distinct forms of surface antigen gene rearrangement in *Trypanosoma brucei*. *Nature*, 297, pp.514–516.
- Majiwa, P.A.O. & Webster, P., 1987. A repetitive deoxyribonucleic acid sequence distinguishes *Trypanosoma simiae* from *T. congolense*. *Parasitology*, 95(3), p.543.
- Malele, I. et al., 2003. The use of specific and generic primers to identify trypanosome infections of wild tsetse flies in Tanzania by PCR. *Infection, Genetics and Evolution*, 3(4), pp.271–279.
- Malvy, D. & Chappuis, F., 2011. Sleeping sickness. *Clinical microbiology and infection: the official publication of the European Society of Clinical Microbiology and Infectious Diseases*, 17(7), pp.986–95.
- Mansfield, J.M. & Paulnock, D.M., 2005. Regulation of innate and acquired immunity in African trypanosomiasis. *Parasite Immunology*, 27(10–11), pp.361–371.
- Marcello, L. & Barry, J.D., 2007a. Analysis of the VSG gene silent archive in *Trypanosoma brucei* reveals that mosaic gene expression is prominent in antigenic variation and is favored by archive substructure. *Genome Research*, 17(9), pp.1344–1352.
- Marcello, L. & Barry, J.D., 2007b. From silent genes to noisy populations - Dialogue between the genotype and phenotypes of antigenic variation. In *Journal of Eukaryotic Microbiology*. pp. 14–17.
- Marchat, L.A. et al., 2015. DEAD/DEXH-Box RNA helicases in selected human parasites. *Korean Journal of Parasitology*, 53(5), pp.583–595.
- Masake, R.A. et al., 1997. Sensitive and specific detection of *Trypanosoma vivax* using the polymerase chain reaction. *Experimental Parasitology*, 85, pp.193–205.
- Masterson, W.J., Taylor, D. & Turner, M.J., 1988. Topologic analysis of the epitopes of a variant surface glycoprotein of *Trypanosoma brucei*. *Journal of Immunology*, 140(9), pp.3194–9.
- Masumu, J. et al., 2006. Comparison of the virulence of *Trypanosoma congolense* strains isolated from cattle in a trypanosomiasis endemic area of eastern Zambia. *International Journal for Parasitology*, 36(4), pp.497–501.
- McCulloch, R. & Barry, J.D., 1999. A role for RAD51 and homologous recombination in *Trypanosoma brucei* antigenic variation. *Genes and Development*, 13(21), pp.2875–2888.
- McCutchan, F.E. et al., 1996. Diversity of the envelope glycoprotein among human immunodeficiency virus type 1 isolates of clade E from Asia and Africa. *Journal of Virology*, 70(6), pp.3331–3338.
- McHardy, A.C. & Adams, B., 2009. The role of genomics in tracking the evolution of influenza A virus. *PLoS Pathogens*, 5(10), p.e1000566.
- McNeillage, G.J.C., Herbert, W.J. & Lumsden, W.H.R., 1969. Antigenic Type of First Relapse Variant Arising from a Strain of *Trypanosoma* (Trypanozoon) *brucei*. *Experimental Parasitology*, 25, pp.1–7.

- Mehlert, A., Bond, C.S. & Ferguson, M. a J., 2002. The glycoforms of a *Trypanosoma brucei* variant surface glycoprotein and molecular modeling of a glycosylated surface coat. *Glycobiology*, 12(10), pp.607–12.
- Van Meirvenne, N. et al., 1975. Antigenic Variation in Syringe Passaged Populations of *Trypanosoma* (Trypanozoon) *brucei*. *Annales De La Societe Belge De Medecine Tropicale*, 55(1), pp.25–30.
- Meirvenne, N. Van, Magnus, E. & Vervoort, T., 1977. Comparisons of variable antigenic types produced by trypanosome strains of the subgenus *Trypanozoon*. *Annual Society Belge Medicine Tropical*, 57(4–5), pp.409–423.
- Melaku, A. & Birasa, B., 2013. Drugs and Drug Resistance in African Animal Trypanosomosis: A Review. *European Journal of Applied Sciences*, 5(3), pp.84–91.
- Melville, S.E. et al., 1998. The molecular karyotype of the megabase chromosomes of *Trypanosoma brucei* and the assignment of chromosome markers. *Molecular and Biochemical Parasitology*, 94(2), pp.155–173.
- Melville, S.E. et al., 2000. The molecular karyotype of the megabase chromosomes of *Trypanosoma brucei* stock 427. *Molecular and Biochemical Parasitology*, 111, pp.261–273.
- Mendoza-Palomares, C. et al., 2008. Molecular and biochemical characterization of a cathepsin B-like protease family unique to *Trypanosoma congolense*. *Eukaryotic Cell*, 7(4), pp.684–697.
- Messenger, L.A., Miles, M.A. & Bern, C., 2015. Between a bug and a hard place: *Trypanosoma cruzi* genetic diversity and the clinical outcomes of Chagas disease. *Expert Review of Anti-Infective Therapy*, 13(8), pp.995–1029.
- Miller, E. & Turner, M., 1981. Analysis of antigenic types appearing in first relapse populations of clones of *Trypanosoma brucei*. *Parasitology*, 82, pp.63–80.
- Miller, E.N., Allan, L.M. & Turner, M.J., 1984a. Mapping of antigenic determinants within peptides of a variant surface glycoprotein of *Trypanosoma brucei*. *Molecular and Biochemical Parasitology*, 13(3), pp.309–322.
- Miller, E.N., Allan, L.M. & Turner, M.J., 1984b. Topological analysis of antigenic determinants on a variant surface glycoprotein of *Trypanosoma brucei*. *Molecular and Biochemical Parasitology*, 13(1), pp.67–81.
- Missiakas, D. & Schneewind, O., 2016. *Staphylococcus aureus* vaccines: Deviating from the carol. *The Journal of Experimental Medicine*, 213(9), pp.1645–1653.
- Moloo, S.K., 1971. An artificial feeding technique for *Glossina*. *Parasitology*, 63(3), pp.507–512.
- Monzon, C.M. et al., 2018. *Trypanosoma vivax* in Argentina. First description. Conferência da Revista Veterinária Argentina.
- Morissette, E. et al., 2009. Diversity of *Anaplasma phagocytophilum* strains, USA. *Emerging Infectious Diseases*, 15(6), pp.928–931.
- Morlais, I. et al., 2001. New molecular marker for *Trypanosoma* (Duttonella) *vivax*

- identification. *Acta Tropica*, 80(3), pp.207–213.
- Morrison, L.J. et al., 2016. Animal African Trypanosomiasis: Time to Increase Focus on Clinically Relevant Parasite and Host Species. *Trends in Parasitology*, 32(8), pp.599–607.
- Morrison, L.J., Tweedie, A., et al., 2009. Discovery of mating in the major African livestock pathogen *Trypanosoma congolense*. *PLoS ONE*, 4(5), p.e5564.
- Morrison, L.J. et al., 2005. Probabilistic order in antigenic variation of *Trypanosoma brucei*. *International Journal for Parasitology*, 35(9), pp.961–972.
- Morrison, L.J. et al., 2010. Role for parasite genetic diversity in differential host responses to *Trypanosoma brucei* infection. *Infection and Immunity*, 78(3), pp.1096–1108.
- Morrison, L.J., Marcello, L. & McCulloch, R., 2009. Antigenic variation in the African trypanosome: Molecular mechanisms and phenotypic complexity. *Cellular Microbiology*, 11(12), pp.1724–1734.
- Moser, D.R. et al., 1989. Detection of *Trypanosoma congolense* and *Trypanosoma brucei* subspecies by DNA amplification using the polymerase chain reaction. *Parasitology*, 99(1), p.57.
- Mossaad, E. et al., 2017. *Trypanosoma vivax* is the second leading cause of camel trypanosomosis in Sudan after *Trypanosoma evansi*. *Parasites & Vectors*, 10(1), p.176.
- Mu, J. et al., 2005. Recombination Hotspots and Population Structure in *Plasmodium falciparum*. *PLoS Biology*, 3(10), p.e335.
- Mugnier, M.R., Cross, G.A.M. & Papavasiliou, F.N., 2015. The in vivo dynamics of antigenic variation in *Trypanosoma brucei*. *Science*, 347(6229), pp.1470–1473.
- Mugnier, M.R., Stebbins, C.E. & Papavasiliou, F.N., 2016. Masters of Disguise: Antigenic Variation and the VSG Coat in *Trypanosoma brucei*. *PLoS Pathogens*, 12(9), p.e1005784.
- Muñoz-Jordán, J.L., Davies, K.P. & Cross, G. a, 1996. Stable expression of mosaic coats of variant surface glycoproteins in *Trypanosoma brucei*. *Science*, 272(May), pp.1795–1797.
- Murray, A.K. & Clarkson, M.J., 1982. Characterization of stocks of *Trypanosoma vivax*. II. Immunological studies. *Annals of Tropical Medicine and Parasitology*, 76, pp.283–292.
- Murrell, B. et al., 2013. FUBAR: A fast, unconstrained bayesian AppRoximation for inferring selection. *Molecular Biology and Evolution*, 30(5), pp.1196–1205.
- Myler, P.J., 1993. Molecular variation in trypanosomes. *Acta Tropica*, 53(3–4), pp.205–225.
- Nacu, S. et al., 2007. Gene expression network analysis and applications to immunology. *Bioinformatics*, 23(7), pp.850–858.
- Namangala, B., Baetselier, P. De, et al., 2000. Attenuation of *Trypanosoma brucei* Is Associated with Reduced Immunosuppression and Concomitant Production of Th2 Lymphokines. *Journal of Infectious Diseases*, 181, pp.1110–1120.
- Namangala, B., 2011. How the African trypanosomes evade host immune killing. *Parasite Immunology*, 33(8), pp.430–437.

- Namangala, B., Brys, L., et al., 2000. *Trypanosoma brucei brucei* infection impairs MHC class II antigen presentation capacity of macrophages. *Parasite Immunology*, 22(7), pp.361–370.
- Nantulya, V.M., Musoke, A.J. & Moloo, S.K., 1986. Apparent exhaustion of the variable antigen repertoires of *Trypanosoma vivax* in infected cattle. *Infection and Immunity*, 54(2), pp.444–447.
- Narayanan, M.S. et al., 2011. NLP is a novel transcription regulator involved in VSG expression site control in *Trypanosoma brucei*. *Nucleic Acids Research*, 39(6), pp.2018–2031.
- Navarro, M. & Gull, K., 2001. A pol I transcriptional body associated with VSG mono-allelic expression in *Trypanosoma brucei*. *Nature*, 414(6865), pp.759–763.
- Ng, S.B. et al., 2009. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*, 461(7261), pp.272–276.
- Njiru, Z.K. et al., 2004. Identification of trypanosomes in *Glossina pallidipes* and *G. longipennis* in Kenya. *Infection, Genetics and Evolution*, 4(1), pp.29–35.
- Ogwu, D. & Nuru, S., 1981. Transplacental transmission of trypanosomes in animals and man. *Veterinary Bulletin*, 51, pp.381–384.
- Okomo-Assoumou, M.C. et al., 1995. Correlation of high serum levels of tumor necrosis factor-alpha with disease severity in human African trypanosomiasis. *American Journal of Tropical Medicine and Hygiene*, 53(5), pp.539–543.
- Oliveira, J.B. et al., 2009. First report of *Trypanosoma vivax* infection in dairy cattle from Costa Rica. *Veterinary Parasitology*, 163(1–2), pp.136–139.
- Ooi, C.-P. et al., 2016. The Cyclical Development of *Trypanosoma vivax* in the Tsetse Fly Involves an Asymmetric Division. *Frontiers in Cellular and Infection Microbiology*, 6(September), pp.1–16.
- Ortega-Montalvo, H.A. et al., 2014. *First report and molecular identification of Trypanosoma vivax in cattle from Ecuador*. Conferência: XIII Congresso Internacional de Parasitologia (ICOPA XIII). Volumen: Resumen 1936. México.
- Osório, A.L.A.R. et al., 2008. *Trypanosoma* (Duttonella) *vivax*: Its biology, epidemiology, pathogenesis, and introduction in the New World - A review. *Memorias do Instituto Oswaldo Cruz*, 103(1), pp.1–13.
- Otto, C., Stadler, P.F. & Hoffmann, S., 2014. Lacking alignments? The next-generation sequencing mapper segemehl revisited. *Bioinformatics*, 30(13), pp.1837–1843.
- Otto, T.D. et al., 2011. RATT: Rapid Annotation Transfer Tool. *Nucleic Acids Research*, 39(9).
- Oyola, S.O. et al., 2016. Whole genome sequencing of *Plasmodium falciparum* from dried blood spots using selective whole genome amplification. *Malaria Journal*, 15(1).
- Paiva, F. et al., 2000. *Trypanosoma vivax* Em Bovinos No Pantanal Do Estado Do Mato Grosso Do Sul , Brasil: I – Acompanhamento Clínico. *Revista Brasileira de Parasitologia Veterinária*, 9(2), pp.135–141.
- Patel, J.C. et al., 2017. Increased risk of low birth weight in women with placental malaria

- associated with *P. falciparum* VAR2CSA clade. *Scientific Reports*, 7(1), p.7768.
- Pays, E., 1989. Pseudogenes, chimaeric genes and the timing of antigen variation in African trypanosomes. *Trends in Genetics*, 5(C), pp.389–391.
- Pays, E., Staerz, U., et al., 1985. Telomeric reciprocal recombination as a possible mechanism for antigenic variation. *Nature*, 316(6028), pp.562–4.
- Pays, E. et al., 1989. The genes and transcripts of an antigen gene expression site from *T. brucei*. *Cell*, 57(5), pp.835–845.
- Pays, E., 2006. The variant surface glycoprotein as a tool for adaptation in African trypanosomes. *Microbes and Infection*, 8(3), pp.930–937.
- Pays, E., Houard, S., et al., 1985. *Trypanosoma brucei*: The extent of conversion in antigen genes may be related to the DNA coding specificity. *Cell*, 42(3), pp.821–829.
- Peacock, C.S. et al., 2008. Comparative genomic analysis of three *Leishmania* species that cause diverse human disease. *Genome*, 39(7), pp.839–847.
- Peacock, L. et al., 2011. Identification of the meiotic life cycle stage of *Trypanosoma brucei* in the tsetse fly. *Proceedings of the National Academy of Sciences*, 108(9), pp.3671–3676.
- Peacock, L. et al., 2009. Intracloal mating occurs during tsetse transmission of *Trypanosoma brucei*. *Parasites & Vectors*, 2(1), p.43.
- Peacock, L., Ferris, V., et al., 2014. Mating compatibility in the parasitic protist *Trypanosoma brucei*. *Parasites & Vectors*, 7(1), p.78.
- Peacock, L., Bailey, M., et al., 2014. Meiosis and haploid gametes in the pathogen *Trypanosoma brucei*. *Current Biology*, 24(2), pp.181–186.
- Peacock, L. et al., 2012. The life cycle of *Trypanosoma* (Nannomonas) *congolense* in the tsetse fly. *Parasites & Vectors*, 5(1), p.109.
- Pearson, W.R., 2013. An introduction to sequence similarity (“homology”) searching. *Current Protocols in Bioinformatics*, 0(3), SUPPL.42.
- Pedram, M. & Donelson, J.E., 1999. The anatomy and transcription of a monocistronic expression site for a metacyclic variant surface glycoprotein gene in *Trypanosoma brucei*. *Journal of Biological Chemistry*, 274(24), pp.16876–16883.
- Peel, E., 1962. Identification of metacyclic trypanosomes in the hypopharynx of tsetse flies, infected in nature or in the laboratory. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 56(4), pp.339–341.
- Pena, A.C. et al., 2014. *Trypanosoma brucei* histone H1 inhibits RNA polymerase I transcription and is important for parasite fitness in vivo. *Molecular Microbiology*, 93(4), pp.645–663.
- Peregrine, A.S., Gray, M.A. & Moloo, S.K., 1997. Cross-Resistance Associated with Development of Resistance to Isometamidium in a Clone of *Trypanosoma congolense*. *Antimicrobial Agents and Chemotherapy*, 41(7), pp.1604–1606.
- Pérez-Morga, D. et al., 2001. Organization of telomeres during the cell and life cycles of *Trypanosoma brucei*. *Journal of Eukaryotic Microbiology*, 48(2), pp.221–226.
- Pillay, D. et al., 2013. *Trypanosoma vivax* GM6 antigen: a candidate antigen for diagnosis of

- African animal trypanosomosis in cattle. *PloS ONE*, 8(10), p.e78565.
- Pinder, M., van Melick, A. & Vernet, G., 1987. Analysis of protective epitopes on the variant surface glycoprotein of a *Trypanosoma brucei brucei* (DiTat 1. 3.) using monoclonal antibodies. *Parasite Immunology*, 9(3), pp.395–400.
- Van der Ploeg, L.H. et al., 1984. Chromosomes of kinetoplastida. *The EMBO journal*, 3(13), pp.3109–3115.
- Pond, S.L.K. & Frost, S.D.W., 2005. Not so different after all: a comparison of methods for detecting amino acid sites under selection. *Molecular biology and evolution*, 22(5), pp.1208–1222.
- Povelones, M.L. et al., 2012. Histone H1 Plays a Role in Heterochromatin Formation and VSG Expression Site Silencing in *Trypanosoma brucei*. *PLoS Pathogens*, 8(11), p.e1003010.
- Proudfoot, C. & McCulloch, R., 2005. Distinct roles for two RAD51-related genes in *Trypanosoma brucei* antigenic variation. *Nucleic Acids Research*, 33(21), pp.6906–6919.
- Purandare, S.M. & Patel, P.I., 1997. Recombination hot spots and human disease. *Genome Research*, 7(8), pp.773–786.
- Quispe, P. et al., 2003. Prevalencia de *Trypanosoma vivax* en bovinos de la provincia de coronel portillo, Ucayali. *Revista de Investigaciones Veterinarias del Perú*, 14(2), pp.161–165.
- Radwanska, M. et al., 2002. The serum resistance-associated gene as a diagnostic tool for the detection of *Trypanosoma brucei rhodesiense*. *American Journal of Tropical Medicine and Hygiene*, 67(6), pp.684–690.
- Radwanska, M. et al., 2008. Trypanosomiasis-Induced B Cell Apoptosis Results in Loss of Protective Anti-Parasite Antibody Responses and Abolishment of Vaccine-Induced Memory Responses. *PLoS Pathogens*, 4(5), p.e1000078.
- Ramey-Butler, K. et al., 2015. Synchronous expression of individual metacyclic variant surface glycoprotein genes in *Trypanosoma brucei*. *Molecular and Biochemical Parasitology*, 200(1–2), pp.1–4.
- Rausch, S. et al., 1994. Sequence determination of three variable surface glycoproteins from *Trypanosoma congolense*. Conserved sequence and structural motifs. *European Journal of Biochemistry / FEBS*, 223(3), pp.813–21.
- Raviprakash, K. et al., 2000. Dengue virus type 1 DNA vaccine induces protective immune responses in rhesus macaques. *Journal of General Virology*, 81(Pt 7), pp.1659–1667.
- Raviprakash, K. et al., 2001. Synergistic neutralizing antibody response to a dengue virus type 2 DNA vaccine by incorporation of lysosome-associated membrane protein sequences and use of plasmid expressing GM-CSF. *Virology*, 290(1), pp.74–82.
- Le Ray, D., Barry, J.D. & Vickerman, K., 1978. Antigenic heterogeneity of metacyclic forms of *Trypanosoma brucei*. *Nature*, 273, pp.300–302.
- Respuela, P. et al., 2008. Histone acetylation and methylation at sites initiating divergent

- polycistronic transcription in *Trypanosoma cruzi*. *Journal of Biological Chemistry*, 283(23), pp.15884–15892.
- Reynolds, D. et al., 2016. Histone H3 Variant Regulates RNA Polymerase II Transcription Termination and Dual Strand Transcription of siRNA Loci in *Trypanosoma brucei*. *PLoS Genetics*, 12(1), p.e1005758.
- Rhoads, A. & Au, K.F., 2015. PacBio Sequencing and Its Applications. *Genomics, Proteomics and Bioinformatics*, 13(5), pp.278–289.
- Ribeiro, J.M.C. et al., 2007. An annotated catalogue of salivary gland transcripts in the adult female mosquito, *Aedes aegypti*. *BMC Genomics*, 8, p.6.
- Ricchetti, M., Dujon, B. & Fairhead, C., 2003. Distance from the chromosome end determines the efficiency of double strand break repair in subtelomeres of haploid yeast. *Journal of Molecular Biology*, 328(4), pp.847–862.
- Robinson, N.P. et al., 1999. Predominance of duplicative VSG gene conversion in antigenic variation in African trypanosomes. *Molecular and Cellular Biology*, 19(9), pp.5839–46.
- Roditi, I. & Lehane, M.J., 2008. Interactions between trypanosomes and tsetse flies. *Current Opinion in Microbiology*, 11(4), pp.345–351.
- Rodrigues, a C. et al., 2008. Phylogenetic analysis of *Trypanosoma vivax* supports the separation of South American/West African from East African isolates and a new *T. vivax*-like genotype infecting a nyala antelope from Mozambique. *Parasitology*, 135(11), pp.1317–1328.
- Rodrigues, C.M. et al., 2017. New insights from Gorongosa National Park and Niassa National Reserve of Mozambique increasing the genetic diversity of *Trypanosoma vivax* and *Trypanosoma vivax*-like in tsetse flies, wild ungulates and livestock from East Africa. *Parasites & Vectors*, 10(1), p.337.
- Ronquist, F. & Huelsenbeck, J.P., 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, 19(12), pp.1572–1574.
- Ross, C.A., Cardoso de Almeida, M.L. & Turner, M.J., 1987. Variant surface glycoproteins of *Trypanosoma congolense* bloodstream and metacyclic forms are anchored by a glycolipid tail. *Molecular and Biochemical Parasitology*, 22(2–3), pp.153–158.
- Rowe, J.A. et al., 2002. Identification of a Conserved *Plasmodium falciparum* var Gene Implicated in Malaria in Pregnancy. *The Journal of Infectious Diseases*, 185(8), pp.1207–1211.
- RStudio Team, -, 2016. RStudio: Integrated Development for R. [Online] RStudio, Inc., Boston, MA, <http://www.rstudio.com>.
- Russell, C.A. et al., 2008. The global circulation of seasonal influenza A (H3N2) viruses. *Science*, 320(5874), pp.340–6.
- Rutherford, K. et al., 2000. Artemis: sequence visualization and annotation. *Bioinformatics*, 16(10), pp.944–945.
- Rutledge, G.G. & Ariani, C. V., 2017. Finding the needle in the haystack. *Nature Reviews Microbiology*, 15(3), p.136.

- Safonova, Y. et al., 2015. Ig Repertoire Constructor: A novel algorithm for antibody repertoire construction and immunoproteogenomics analysis. *Bioinformatics*, 31(12), pp.i53–i61.
- Salanti, A. et al., 2003. Selective upregulation of a single distinctly structured var gene in chondroitin sulphate A-adhering *Plasmodium falciparum* involved in pregnancy-associated malaria. *Molecular Microbiology*, 49(1), pp.179–191.
- Salmon, D. et al., 1994. A novel heterodimeric transferrin receptor encoded by a pair of VSG expression site-associated genes in *T. brucei*. *Cell*, 78(1), pp.75–86.
- Salmon, D. et al., 2012. Adenylate cyclases of *Trypanosoma brucei* inhibit the innate immune response of the host. *Science*, 337(6093), pp.463–466.
- Salmon, D. et al., 1997. Characterization of the ligand-binding site of the transferrin receptor in *Trypanosoma brucei* demonstrates a structural relationship with the N-terminal domain of the variant surface glycoprotein. *The EMBO Journal*, 16(24), pp.7272–7278.
- Savage, A.F. et al., 2016. Transcriptome Profiling of *Trypanosoma brucei* Development in the Tsetse Fly Vector *Glossina morsitans*. *PLoS ONE*, 11(12), p.e0168877.
- Schell, D. et al., 1991. A transferrin-binding protein of *Trypanosoma brucei* is encoded by one of the genes in the variant surface glycoprotein gene expression site. *EMBO Journal*, 10(5), pp.1061–1066.
- Schleifer, K.W. et al., 1993. Characterization of T helper cell responses to the trypanosome variant surface glycoprotein. *Journal of Immunology*, 150(7), pp.2910–2919.
- Schulz, D. et al., 2016. Base J and H3.V Regulate Transcriptional Termination in *Trypanosoma brucei*. *PLoS Genetics*, 12(1), p.e1005762.
- Schwede, A. et al., 2015. How Does the VSG Coat of Bloodstream Form African Trypanosomes Interact with External Proteins? *PLoS Pathogens*, 11(12), p.e1005259.
- Sendashonga, C.N. & Black, S.J., 1982. Humoral responses against *Trypanosoma brucei* variable surface antigen are induced by degenerating parasites. *Parasite Immunology*, 4, pp.245–257.
- Shaw, J.J. & Lainson, R., 1972. *Trypanosoma vivax* in Brazil. *Annals of Tropical Medicine and Parasitology*, 66(1), pp.25–32.
- Shea, C., Lee, M.G.S. & Van der Ploeg, L.H.T., 1987. VSG gene 118 is transcribed from a cotransposed pol I-like promoter. *Cell*, 50(4), pp.603–612.
- Shea, C. & Van der Ploeg, L.H., 1988. Stable variant-specific transcripts of the variant cell surface glycoprotein gene 1.8 expression site in *Trypanosoma brucei*. *Molecular and Cell Biology*, 8(2), pp.854–859.
- Shi, M., Pan, W. & Tabel, H., 2003. Experimental African trypanosomiasis: IFN-gamma mediates early mortality. *European Journal of Immunology*, 33(1), pp.108–118.
- Shimodaira, H. & Hasegawa, M., 1999. Multiple Comparisons of Log-Likelihoods with Applications to Phylogenetic Inference. *Molecular Biology and Evolution*, 16(8), pp.1114–1116.
- Siegel, T.N. et al., 2010. Genome-wide analysis of mRNA abundance in two life-cycle stages of *Trypanosoma brucei* and identification of splicing and polyadenylation sites. *Nucleic*

- Acids Research*, 38(15), pp.4946–4957.
- Sigauque, I. et al., 2000. The distribution of tsetse (Diptera: *Glossinidae*) and bovine trypanosomosis in the Matutuine District, Maputo Province, Mozambique. *The Onderstepoort Journal of Veterinary Research*, 67, pp.167–172.
- Sileghem, M. et al., 1994. Tumour necrosis factor production by monocytes from cattle infected with *Trypanosoma* (Duttonella) *vivax* and *Trypanosoma* (Nannomonas) *congolense*: possible association with severity of anaemia associated with the disease. *Parasite Immunology*, 16(1), pp.51–54.
- Silva, R.A. et al., 1996. Outbreak of trypanosomiasis due to *Trypanosoma vivax* in bovines of the Pantanal, Brazil. *Memórias do Instituto Oswaldo Cruz*, 91(5), pp.561–562.
- Silva, R.A.M.S. et al., 1998. Outbreaks of trypanosomosis due to *Trypanosoma vivax* in cattle in Bolivia. *Veterinary Parasitology*, 76(1–2), pp.153–157.
- Silvester, E. et al., 2017. Interspecies quorum sensing in co-infections can manipulate trypanosome transmission potential. *Nature Microbiology*, 2, pp. 1471–1479.
- Simo, G. et al., 2013. Identification and genetic characterization of *Trypanosoma congolense* in domestic animals of Fontem in the South-West region of Cameroon. *Infection, Genetics and Evolution*, 18, pp.66–73.
- Simo, G., Fogue, P., Melachio, T.T., et al., 2014. Population genetics of forest type of *Trypanosoma congolense* circulating in *Glossina palpalis palpalis* of Fontem in the South-West region of Cameroon. *Parasites & Vectors*, 7(1), p.385.
- Simo, G., Fogue, P.S., Melachio, T.T.T., et al., 2014. Population genetics of forest type of *Trypanosoma congolense* circulating in *Glossina palpalis palpalis* of Fontem in the South-West region of Cameroon. *Parasites & Vectors*, 7, p.385.
- Simpson, J.T. et al., 2009. ABySS: A parallel assembler for short read sequence data. *Genome Research*, 19(6), pp.1117–1123.
- Sistrom, M. et al., 2014. Comparative genomics reveals multiple genetic backgrounds of human pathogenicity in the *Trypanosoma brucei* complex. *Genome Biology and Evolution*, 6(10), pp.2811–2819.
- Sloof, P. et al., 1983. Characterization of satellite DNA in *Trypanosoma brucei* and *Trypanosoma cruzi*. *Journal of Molecular Biology*, 167(1), pp.1–21.
- Smetko, A. et al., 2015. Trypanosomosis: potential driver of selection in African cattle. *Frontiers in Genetics*, 6(April), pp.1–8.
- Smith, D.J. et al., 2004. Mapping the antigenic and genetic evolution of influenza virus. *Science*, 305(5682), pp.371–376.
- Soltys, M.A., 1963. Immunity in African trypanosomiasis. *Bulletin of the World Health Organization*, 28(1952), pp.753–761.
- de Souza Pimentel, D. et al., 2012. First report and molecular characterization of *Trypanosoma vivax* in cattle from state of Pernambuco, Brazil. *Veterinary Parasitology*, 185(2–4), pp.286–289.
- Spence, P.J. et al., 2013. Vector transmission regulates immune control of *Plasmodium*

- virulence. *Nature*, 498, pp.228–31.
- Stafford-Banks, C.A. et al., 2014. Analysis of the Salivary Gland Transcriptome of *Frankliniella occidentalis*. *PLoS ONE*, 9(4), p.e94447.
- Stamatakis, A., 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9), pp.1312–1313.
- Stanke, M. et al., 2004. AUGUSTUS: A web server for gene finding in eukaryotes. *Nucleic Acids Research*, 32(WEB SERVER ISS.), pp. W309–W312.
- Stanne, T. et al., 2015. Identification of the ISWI chromatin remodeling complex of the early branching eukaryote *Trypanosoma brucei*. *Journal of Biological Chemistry*, 290(45), pp.26954–26967.
- Steinbiss, S. et al., 2016. *Companion*: a web server for annotation and analysis of parasite genomes. *Nucleic Acids Research*, 44(11), p.gkw292.
- Stephen, L.E., 1986. *Trypanosomiasis: a veterinary perspective*. Oxford: Pergamon Press.
- Steverding, D., 2008. The history of African trypanosomiasis. *Parasites & Vectors*, 1(1), p.3.
- Stewart, M.L. et al., 2010. Multiple genetic mechanisms lead to loss of functional TbAT1 expression in drug-resistant trypanosomes. *Eukaryotic Cell*, 9(2), pp.336–343.
- Stijlemans, B. et al., 2007. African trypanosomiasis: From immune escape and immunopathology to immune intervention. *Veterinary Parasitology*, 148(1 SPEC. ISS.), pp.3–13.
- Stijlemans, B. et al., 2016. Immune evasion strategies of *Trypanosoma brucei* within the mammalian host: Progression to pathogenicity. *Frontiers in Immunology*, 7(JUN), pp.233.
- Stijlemans, B. et al., 2008. Role of iron homeostasis in trypanosomiasis-associated anemia. *Immunobiology*, 213(9–10), pp.823–835.
- Stijlemans, B. et al., 2010. The central role of macrophages in trypanosomiasis-associated anemia: rationale for therapeutical approaches. *Endocrine, Metabolic & Immune Disorders Drug Targets*, 10(1), pp.71–82.
- Strahl, B.D. & Allis, C.D., 2000. The language of covalent histone modifications. *Nature*, 403(6765), pp.41–45.
- Stranger-Jones, Y.K., Bae, T. & Schneewind, O., 2006. Vaccine assembly from surface proteins of *Staphylococcus aureus*. *Proceedings of the National Academy of Sciences*, 103, pp.16942–16947.
- Strickler, J.E. et al., 1987. *Trypanosoma congolense*: Structure and Molecular Organization of the Surface Glycoproteins of Two Early Bloodstream Variants. *Biochemistry*, 26, pp.796–805.
- Stuart, K. et al., 2008. Review series Kinetoplastids : related protozoan pathogens , different diseases. *The Journal of Clinical Investigation*, 118(4), pp.1301–1310.
- Sunter, J., Webb, H. & Carrington, M., 2013. Determinants of GPI-PLC Localisation to the Flagellum and Access to GPI-Anchored Substrates in Trypanosomes. *PLoS Pathogens*, 9(8), p.e1003566.

- Swallow, B., 1999. Impacts of trypanosomiasis on African agriculture. *International Livestock Research Institute, Nairobi, Kenya.*, pp.1–46.
- Tachado, S.D. & Schofield, L., 1994. Glycosylphosphatidylinositol toxin of *Trypanosoma brucei* regulates IL-1 alpha and TNF-alpha expression in macrophages by protein tyrosine kinase mediated signal transduction. *Biochemical and Biophysical Research Communications*, 205(2), pp.984–91.
- Tait, A. et al., 2007. Genetic exchange in *Trypanosoma brucei*: Evidence for mating prior to metacyclic stage development. *Molecular and Biochemical Parasitology*, 151(1), pp.133–136.
- Tait, A. & Turner, C.M., 1990. Genetic exchange in *Trypanosoma brucei*. *Parasitology Today*, 6(3), pp.70–5.
- Tamura, K. & Nei, M., 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular Biology and Evolution*, 10(3), pp.512–26.
- Tanowitz, H.B. et al., 2017. Adipose Tissue: A Safe Haven for Parasites? *Trends in Parasitology*, 33(4), pp.276–284.
- Tarique, M. et al., 2013. *Plasmodium falciparum* DOZI, an RNA helicase interacts with eIF4E. *Gene*, 522(1), pp.46–59.
- Taylor, H.M., Kyes, S.A. & Newbold, C.I., 2000. Var gene diversity in *Plasmodium falciparum* is generated by frequent recombination events. *Molecular and Biochemical Parasitology*, 110(2), pp.391–397.
- Tetley, L. et al., 1987. Onset of expression of the variant surface glycoproteins of *Trypanosoma brucei* in the tsetse fly studied using immunoelectron microscopy. *Journal of Cell Science*, 87 (Pt 2)(2), pp.363–72.
- Thon, G., Baltz, T. & Eisen, H., 1989. Antigenic diversity by the recombination of pseudogenes. *Genes & Development*, 3(8), pp.1247–1254.
- Tihon, E. et al., 2017. Discovery and genomic analyses of hybridization between divergent lineages of *Trypanosoma congolense*, causative agent of Animal African Trypanosomiasis. *Molecular Ecology*, 6(23), pp. 6524–6538.
- Timmers, H.T.M. et al., 1987. Coincident multiple activations of the same surface antigen gene in *Trypanosoma brucei*. *Journal of Molecular Biology*, 194(1), pp.81–90.
- Trapnell, C. et al., 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature protocols*, 7(3), pp.562–78.
- Trapnell, C. et al., 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, 28(5), pp.511–515.
- Trindade, S. et al., 2016. *Trypanosoma brucei* Parasites Occupy and Functionally Adapt to the Adipose Tissue in Mice. *Cell Host and Microbe*, 19(6), pp.837–848.
- Tuikue Ndam, N. et al., 2008a. *Plasmodium falciparum* Transcriptome Analysis Reveals Pregnancy Malaria Associated Gene Expression. *PLoS ONE*, 3(3), p.e1855.

- Tuikue Ndam, N. et al., 2008b. *Plasmodium falciparum* transcriptome analysis reveals pregnancy malaria associated gene expression. *PLoS ONE*, 3(3), p.e1855.
- Turner, C.M.R. et al., 1988. An estimate of the size of the metacyclic variable antigen repertoire of *Trypanosoma brucei rhodesiense*. *Parasitology*, 97(2), pp.269–276.
- Turner, C.M.R. et al., 1990. Evidence that the mechanism of gene exchange in *Trypanosoma brucei* involves meiosis and syngamy. *Parasitology*, 101(3), pp.377–386.
- Turner, C.M.R., Aslam, N. & Dye, C., 1995. Replication, Differentiation, Growth And The Virulence Of *Trypanosoma brucei* Infections. *Parasitology*, 111, pp.289–300.
- Ukaegbu, U.E. et al., 2015. A Unique Virulence Gene Occupies a Principal Position in Immune Evasion by the Malaria Parasite *Plasmodium falciparum*. *PLoS Genetics*, 11(5), p.e1005234.
- Uzureau, P. et al., 2013. Mechanism of *Trypanosoma brucei gambiense* resistance to human serum. *Nature*, 501(7467), pp.430–4.
- Valdés, J. et al., 2014. Proteomic analysis of *Entamoeba histolytica* in vivo assembled pre-mRNA splicing complexes. *Journal of Proteomics*, 111, pp.30–45.
- Vanhamme, L. et al., 2003. Apolipoprotein L-I is the trypanosome lytic factor of human serum. *Nature*, 422(6927), pp.83–87.
- Vanhamme, L. et al., 2004. The *Trypanosoma brucei* reference strain TREU927/4 contains *T. brucei rhodesiense*-specific SRA sequences, but displays a distinct phenotype of relative resistance to human serum. *Molecular and Biochemical Parasitology*, 135(1), pp.39–47.
- Ventura, R.M. et al., 2001. *Trypanosoma vivax*: characterization of the spliced-leader gene of a Brazilian stock and species-specific detection by PCR amplification of an intergenic spacer sequence. *Experimental Parasitology*, 99(1), pp.37–48.
- Vickerman, K., 1969. On the surface coat and flagellar adhesion in trypanosomes. *Journal of Cell Science*, 5, pp.163–193.
- Vickerman, K. & Preston, T.M., 1970. Spindle microtubules in the dividing nuclei of trypanosomes. *Journal of Cell Science*, 6, pp.365–383.
- Vincendeau, P. & Bouteille, B., 2006. Immunology and immunopathology of African trypanosomiasis. *Anais da Academia Brasileira de Ciências*, 78(4), pp.645–665.
- Virji, M., 2009. Pathogenic *Neisseriae*: surface modulation, pathogenesis and infection control. *Nature reviews. Microbiology*, 7(4), pp.274–86.
- Wang, C.W. et al., 2012. Genetic diversity of expressed *Plasmodium falciparum* var genes from Tanzanian children with severe malaria. *Malaria Journal*, 11(1), p.230.
- Wang, J., Böhme, U. & Cross, G.A.M., 2003. Structural features affecting variant surface glycoprotein expression in *Trypanosoma brucei*. *Molecular and Biochemical Parasitology*, 128(2), pp.135–145.
- Wang, X. et al., 2012. Characterization of the unusual bidirectional ves promoters driving vesa1 expression and associated with antigenic variation. *Eukaryotic Cell*, 11(3), pp.260–269.

- Weir, W. et al., 2016. Population genomics reveals the origin and asexual evolution of human infective trypanosomes. *eLife*, 5(January 2016), p.e11473.
- Wellde, B.T. et al., 1983. Haemorrhagic syndrome in cattle associated with *Trypanosoma vivax* infection. *Tropical Animal Health and Production*, 15(2), pp.95–102.
- Wells, E.A., Ramirez, L.E. & Betancourt, A., 1982. *Trypanosoma vivax* in Colombia: Interpretation of field results. *Tropical Animal Health and Production*, 14(3), pp.141–150.
- Wells, J.M. et al., 1987. DNA contents and molecular karyotypes of hybrid *Trypanosoma brucei*. *Molecular and Biochemical Parasitology*, 24(1), pp.103–116.
- Wickstead, B., Ersfeld, K. & Gull, K., 2002. Targeting of a tetracycline-inducible expression system to the transcriptionally silent minichromosomes of *Trypanosoma brucei*. *Molecular and Biochemical Parasitology*, 125(1–2), pp.211–216.
- Wickstead, B., Ersfeld, K. & Gull, K., 2003. The mitotic stability of the minichromosomes of *Trypanosoma brucei*. *Molecular and Biochemical Parasitology*, 132, pp.97–100.
- Wickstead, B., Ersfeld, K. & Gull, K., 2004. The Small Chromosomes of *Trypanosoma brucei* Involved in Antigenic Variation Are Constructed Around Repetitive Palindromes. *Genome Research*, pp.1014–1024.
- Wilkes, J.M. et al., 1997. Modulation of mitochondrial electrical potential: a candidate mechanism for drug resistance in African trypanosomes. *Biochemical Journal*, 326, pp.755–761.
- Winther-Larsen, H.C. et al., 2005. A conserved set of pilin-like molecules controls type IV pilus dynamics and organelle-associated functions in *Neisseria gonorrhoeae*. *Molecular Microbiology*, 56(4), pp.903–917.
- Woo, P.T.K., 1970. The haematocrit centrifuge technique for the diagnosis of African trypanosomiasis. *Acta Tropica*, 27(4), pp.384–6.
- Van Xong, H. et al., 1998. A VSG expression site-associated gene confers resistance to human serum in *Trypanosoma rhodesiense*. *Cell*, 95(6), pp.839–846.
- Young, C.J. & Godfrey, D.G., 1983. Enzyme polymorphism and the distribution of *Trypanosoma congolense* isolates. *Annals of Tropical Medicine and Parasitology*, 77(5), pp.467–81.
- Young, J.R. et al., 1982. Analysis of genomic rearrangements associated with two variable antigen genes in *Trypanosoma brucei*. *Nucleic Acids Research*, 10(3).
- Young, J.R. et al., 1983. Are there two classes of VSG gene in *Trypanosoma brucei*? *Nature*, 306, pp.196–198.
- Zerbino, D.R., 2010. Using the Velvet de novo assembler for short-read sequencing technologies. *Current Protocols in Bioinformatics*, (SUPPL. 31), Unit 11.5.
- Zerbino, D.R. & Birney, E., 2008. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, 18(5), pp.821–829.
- Zhuang, Y. et al., 2007. Maintenance of antibody to pathogen epitopes generated by segmental gene conversion is highly dynamic during long-term persistent infection. *Infection and Immunity*, 75(11), pp.5185–5190.

- Zingales, B., 2017. *Trypanosoma cruzi* genetic diversity: Something new for something known about Chagas disease manifestations, serodiagnosis and drug sensitivity. *Acta Tropica*, (September), pp.0–1.
- Zinoviev, A. et al., 2011. A novel 4E-interacting protein in *Leishmania* is involved in stage-specific translation pathways. *Nucleic Acids Research*, 39(19), pp.8404–8415.